AD_____

CONTRACT NUMBER DAMD17-94-C-4082

TITLE: Increasing the Accuracy of Mammogram Interpretation

PRINCIPAL INVESTIGATOR: John A. Swets, Ph.D.

CONTRACTING ORGANIZATION: BBN Technologies
GTE Internetworking
Cambridge, Massachusetts 02138

REPORT DATE: October 1998

TYPE OF REPORT: Final

PREPARED FOR: Commander
U.S. Army Medical Research and Materiel Command
Fort Detrick, Frederick, Maryland 21702-5012

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. AGENCY USE ONLY *(Leave blank)*	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
	October 1998	Final (14 Sep 94 - 14 Sep 98)

4. TITLE AND SUBTITLE
Increasing the Accuracy of Mammogram Interpretation

5. FUNDING NUMBERS
DAMD17-94-C-4082

6. AUTHOR(S)
John A. Swets, Ph.D.

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
BBN Technologies
GTE Internetworking
Cambridge, Massachusetts 02138

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)
U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

10. SPONSORING / MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION / AVAILABILITY STATEMENT
Approved for Public Release; Distribution Unlimited

12b. DISTRIBUTION CODE

13. ABSTRACT *(Maximum 200 words)* A computer-based decision-support system and automated report writer for mammography were refined and evaluated in a clinical setting. For each case, the computer solicits from a radiologist quantitative ratings of a checklist of perceptual features of mammograms that have been determined to be most diagnostic and most informative for a therapeutic recommendation. The ratings are converted both to a probability of malignancy by a statistical prediction rule and to a prose report of findings by computational linguistic techniques. Overall, the decision aids served to increase accuracy less than in previous laboratory studies, but a substantial gain was shown for the more difficult cases that present as calcifications. Other substantive and methodological advances show the continuing promise of this approach for further development and use in practice. The version of a report writer that was evaluated was seen to need further development along specifiable lines, but demonstrated its ability to improve on the usual dictated report in several respects. Together, the decision-support system and automated report writer are expected to find cost-effective use in a larger radiological and medical information system. The ranked list of certified, scaled perceptual features developed by this approach, and the predictive value of their merged ratings, should also be valuable as a teaching tool.

14. SUBJECT TERMS
Breast Cancer, mammography, decision-support system, automated report writing, quality assurance

15. NUMBER OF PAGES
80

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
Unclassified	Unclassified	Unclassified	Unlimited

FOREWORD

Opinions, interpretations, conclusions and recommendations are those of the author and are not necessarily endorsed by the U.S. Army.

____ Where copyrighted material is quoted, permission has been obtained to use such material.

____ Where material from documents designated for limited distribution is quoted, permission has been obtained to use the material.

__X__ Citations of commercial organizations and trade names in this report do not constitute an official Department of Army endorsement or approval of the products or services of these organizations.

____ In conducting research using animals, the investigator(s) adhered to the "Guide for the Care and Use of Laboratory Animals," prepared by the Committee on Care and use of Laboratory Animals of the Institute of Laboratory Resources, national Research Council (NIH Publication No. 86-23, Revised 1985).

__X__ For the protection of human subjects, the investigator(s) adhered to policies of applicable Federal Law 45 CFR 46.

____ In conducting research utilizing recombinant DNA technology, the investigator(s) adhered to current guidelines promulgated by the National Institutes of Health.

____ In the conduct of research utilizing recombinant DNA, the investigator(s) adhered to the NIH Guidelines for Research Involving Recombinant DNA Molecules.

____ In the conduct of research involving hazardous organisms, the investigator(s) adhered to the CDC-NIH Guide for Biosafety in Microbiological and Biomedical Laboratories.

John A Sweets 30 Sept '98

PI - Signature Date

TABLE OF CONTENTS

INTRODUCTION

The aims of this project were: (1) to refine a decision-support system for interpreting mammograms and extend its evaluation from a laboratory to clinical setting, and (2) to develop as an addition to the system an automated report writer and give it a formative evaluation.

The decision-support system provides to the radiologist a checklist of the several perceptual features of a mammogram that have been statistically determined to be diagnostically most relevant. It solicits from the radiologist, for a given clinical case, a rating of each feature, usually a value on a 10-point scale. The system merges these scale values optimally via a statistical prediction rule (SPR) to calculate a probability of malignancy that it submits to the radiologist as an advisory. Also, from the pattern of feature values, the system automatically constructs a prose report of findings from the radiologist to the referring physician and surgeon.

Earlier laboratory experiments showed for a previous version of the system that together its two aids to mammogram reading—that is, the checklist of features and the SPR—produced substantial increases in diagnostic accuracy (Getty, Pickett, D'Orsi, and Swets, 1988; Swets, Getty, Pickett, D'Orsi, Seltzer, and McNeil, 1991; D'Orsi, Getty, Swets, Pickett, Seltzer, and McNeil, 1992; Seltzer, McNeil, D'Orsi, Getty, Pickett, and Swets, 1992). The refinement of the system in the present project was undertaken primarily to include observed changes in perceptual features from previous to current mammograms, so-called "interval change." Further, the SPR in the newer version is based on a logistic-regression analysis rather than the linear-discriminant analysis used previously. And our experimental design permitted separating the effects of the checklist and the SPR's advisory probability.

The extension of the system's evaluation from laboratory to clinic was achieved by developing the SPR with the mammogram specialists and cases of a referral, diagnostic center—the Brigham and Women's Hospital (BWH)—and applying it to the general radiologists and cases of a community-based, screening setting—of Harvard Pilgrim Health Care (HPHC). For additional realism in this evaluation, the case sets used were extended beyond those of earlier studies, which contained biopsied cases and cases considered normal but followed for several years, to include patients for whom a suspected abnormality turned out to be a normal variation.

The automated report writer was brought to a level of computational and linguistic sophistication that was appropriate for its formative evaluation. The evaluation was carried out by generating both automated reports and typical, dictated reports at BWH and submitting them for comparative review to a BWH panel composed of clinicians (referring physicians) and surgeons. We discuss our studies of the decision-support system and automated report writer in turn, as parts 1 and 2 of the following "body of report."

BODY OF REPORT

1. Decision-Support System

1.1 Experimental design

Five mammographers at BWH assigned ratings to an exhaustive set of 66 perceptual features for 211 proven BWH cases—both to determine a subset of features necessary and sufficient for diagnosis and to "train" an SPR to estimate the probability of malignancy as based on that subset of feature ratings. Actually, there were two subsets of features and two corresponding SPRs, one for cases presenting as masses and the other for cases presenting as calcifications. To anticipate, each SPR included eight features.

Five radiologists at HPHC then assigned ratings to the features of the appropriate SPR for a set of 150 proven HPHC cases—to "test" the SPR. In order to determine the gain in accuracy provided by the decision-support system, the accuracy of mammogram interpretations at HPHC—made with the aid both of the final list of perceptual features and the probability estimate of malignancy—was compared to a baseline accuracy of the interpretations of the same cases made there a few months earlier by the same radiologists, in their usual manner. A measure of accuracy was also taken based only on the checklist aid, without the probability estimate, to separate the contributions to increased accuracy of the two component aids. As an extension of this basic experimental design, we also developed mass and calcification SPRs based on the feature ratings of the HPHC radiologists, who had the opportunity to work intensively with the final, relatively tractable, set of features developed at BWH.

1.2 Accuracy assessments

Assessments of accuracy were made by ROC analysis (relative, or receiver, operating characteristic) to obtain an index of accuracy that is unaffected by an observer's decision threshold and by the relative frequencies (prior probabilities) of malignant and non-malignant cases in the test set (Swets, 1979; Swets, 1986a, b; 1988; Swets and Pickett, 1982). ROCs – which plot the true-positive proportion (TPP) against the false-positive proportion (FPP) -- were constructed from the estimates of probability of malignancy that were made by radiologist observers and also from probability estimates made by the SPR. The accuracy measure A_z, i.e., the area under the ROC based on normal distributions, was calculated by one of Charles Metz's ensemble of ROC-fitting computer programs, for correlated or uncorrelated data as appropriate (see Swets, 1996). A_z can vary from .50 to 1.0. Because it reflects the locus of a curve based on several empirical points, A_z is determined with a substantially smaller error of estimate than is any single point on the curve, i.e., a single pair of sensitivity and specificity values. The values of A_z shown later reflect the group performances of each set of five radiologists; they are either given as mean values of A_z or as the single A_z obtained from a group ROC that was produced by pooling probability estimates of five radiologists. Statistical tests of differences in A_z values (one-tailed) were also based on Metz's programs. They are conservative tests in that whereas they do take into account that a given case set is read in two conditions, they do not take into account correlation across readers.

1.3 Case selection

Cases were obtained retrospectively at the two clinical sites and were selected to represent categories of cases for which a decision-support system would tend to be used, namely, malignancies, benign lesions, and "suspicious" cases that were

determined subsequently to be "normal." Images taken at two different times were included—the images first deemed suspicious and the images of the last preceding examination—in order to examine interval change. The Human Research Committees of both institutions approved the study protocol. Table I summarizes the eligibility criteria and the enrollment statistics of the final case sets.

Table I. Eligibility Criteria and Enrollment Statistics

Category	Definition	Method of Proof	Final Case Sets at BWH	HPHC
Malignant	All types of breast cancer except lobular carcinoma-in-situ	Pathology	107	50
Benign	Focal, nonmalignant lesions (i.e., benign tumors)	Pathology	53	51
Suspicious	Patient referred for additional imaging studies or accelerated follow-up and not returned to routine screening pool	Clinical/Imaging (i.e., no change in lesion appearance during monitoring)	51	49
			(Total = 211)	(Total = 150)

Table II shows the distribution of cases over types of mammographic presentation. Masses and clustered calcifications are the most common presentations. The distributions of relevant classes of patients selected fit the demographics of the mammography referral (BWH) and screening (HPHC) practices at our two sites. This fit ensures adequate enrollment of minority groups. For each eligible case, all available original mammographic and ultrasound images at the time of the "target" examination (i.e., when the suspicious focus was identified) were harvested for use. In addition, in order to support development and evaluation of interval-change features, mammographic images from a "comparison" examination dating approximately 12 months (range 6 to 18 months) before the target examination were also selected.

Patient-identifying information was covered by removable tape and a study number assigned to each case to ensure patient confidentiality.

Table II. Distribution of Cases over Types of Mammographic Indication

	BWH				HPHC			
TYPE	Suspicious Normal	Benign	Malignant	Total	Suspicious Normal	Benign	Malignant	Total
Architectural distortion	1	1	10	12	1	0	0	1
Asymmetric breast tissue	7	4	3	14	17	1	2	20
Clustered calcifications	5	30	35	70	10	23	17	50
Regional calcifications	3	4	6	13	0	3	0	3
Mass	35	14	53	102	21	24	31	76

The images in each case enrolled were checked for technical quality at BWH by one of the BWH investigators (T.F.) and at HPHC by one of the HPHC investigators (J.M.). In addition, these individuals confirmed that all needed views were available and confirmed selection of the appropriate comparison study. In preparation for training the SPR, investigator T. F. and another of the BWH investigators (J.E.M.) reviewed all the selected images and listed the coordinates of the most suspicious mammographic abnormality. This step ensured that the radiologists rendered feature ratings on the same lesion. At HPHC, the investigator J. M. performed the same review for its cases.

1.4 Selection of features and SPR development

The necessary and sufficient sets of features for the checklist and two SPRs were determined from the BWH ratings by a stepwise, logistic-regression procedure in which features were added one-by-one to the system. At each step, that feature was added that would most improve the SPR, given the set of features already included. Features were added until additional features failed to make a statistically significant contribution (Hosmer and Lemeshow, 1989).

The SPR's estimate of probability of malignancy is derived from a function of the linear, weighted sum of the feature ratings supplied by the radiologist. The feature weights are maximum-likelihood estimates whose magnitudes depend both on the diagnostic contribution of each feature and on the pattern of intercorrelations among features.

1.5 Checklist development

The total list of features to be used by HPHC radiologists was converted into modular form. Specifically, a separate module was created for each of the five different kinds of radiographic presentation of the lesions that occurred in both the BWH and HPHC cases samples: 1) Mass; 2) Not-Definitely-Benign (Clustered) Calcifications; 3) Asymmetric Breast Tissue; 4) Architectural Distortion; and 5) Regional Calcifications. This modular arrangement enabled the radiologist reader to proceed efficiently, following just that module (or, potentially, the combination of modules) that applied to the case at hand. The five modules are shown in Appendix A.

1.6 Format of radiologist input to the system

Our initial plan was to provide for computer recognition of rating values spoken by the radiologist. However, the best speech recognition system we could find at the

time (Phonetic Engine 500) turned out to be not sufficiently robust to use with radiologists who read the films at times of their convenience, alone, without the presence of a project technical assistant. We did use our recognition system to enter data off line in our laboratory that the BWH mammographers had supplied on paper-and-pencil forms and we used our experience with system failures then to arrange system improvements with the manufacturer (Speech Systems Inc). Nonetheless, we deemed the system inadequate for clinical use at the time. We understand that an IBM system available now, and in use at BWH, might be adequate for our original purposes.

1.7 Results: BWH readings

The BWH ratings of 66 perceptual features were the basis for the two SPRs (one for masses and one for calcifications, i.e., for the two subsets of cases of clearly different types that existed in large enough numbers to yield an acceptably reliable SPR for each). We illustrate here the procedure of developing an SPR just for masses. Table III shows 24 candidate features for the mass SPR ranked according to their predictive power (labeled "significance level" in the table) when considered individually, that is, at "Step 0" of the stepwise procedure. The first 18 of the 24 features are seen individually to have a statistically significant predictive value ($p < .01$).

Figure 1 shows for masses how the significance levels of the 18 significant features varied over successive steps in a regression analysis (for a pre-final version of the SPR). The features are laid out along the Z-axis, the step number in the regression along the X-axis, and the feature significance level along the Y-axis. It is seen at Step 0 that all of the features are highly significant predictors. The features that are incorporated into this version of the SPR are all located at the far end of the box. With increasing step number, features at the near end of the box have significance levels that start out as highly significant, but fall off rapidly as other features are incorporated into the rule.

Table III. Candidate Features for Masses

	Mass Feature	Score	Sig. Level
1.	Shape	285.2160	.0000
2.	% Margin Circumscribed	276.7988	.0000
3.	Tissue Invasion	264.7068	.0000
4.	Spiculation	248.1468	.0000
5.	% Margin Invaded	171.1519	.0000
6.	Related Architectural Distortion	155.7087	.0000
7.	% Margin Spiculated	139.5676	.0000
8.	Microlobulation	116.3964	.0000
9.	Intramammary Node	109.0096	.0000
10.	Fat Inclusion	81.5289	.0000
11.	Related Worrisome Calcifications	65.7233	.0000
12.	Age	45.4388	.0000
13.	Density	34.2371	.0000
14.	% Margin Invaded	25.6192	.0000
15.	Related Benign Calcifications	15.1205	.0001
16.	Skin Lesion	14.7566	.0001
17.	Mass Distribution	10.7151	.0001
18.	Minimum Diameter	6.0719	.0137
19.	Change in Glandular Density	3.0109	.0827
20.	Change in % Glandular Tissue	2.6430	.1040
21.	% Glandular Tissue	.9339	.3339
22.	Aspect Ratio	.3603	.5483
23.	% Margin Obscured	.1227	.7261
24.	Density of Glandular Tissue	.0277	.8678

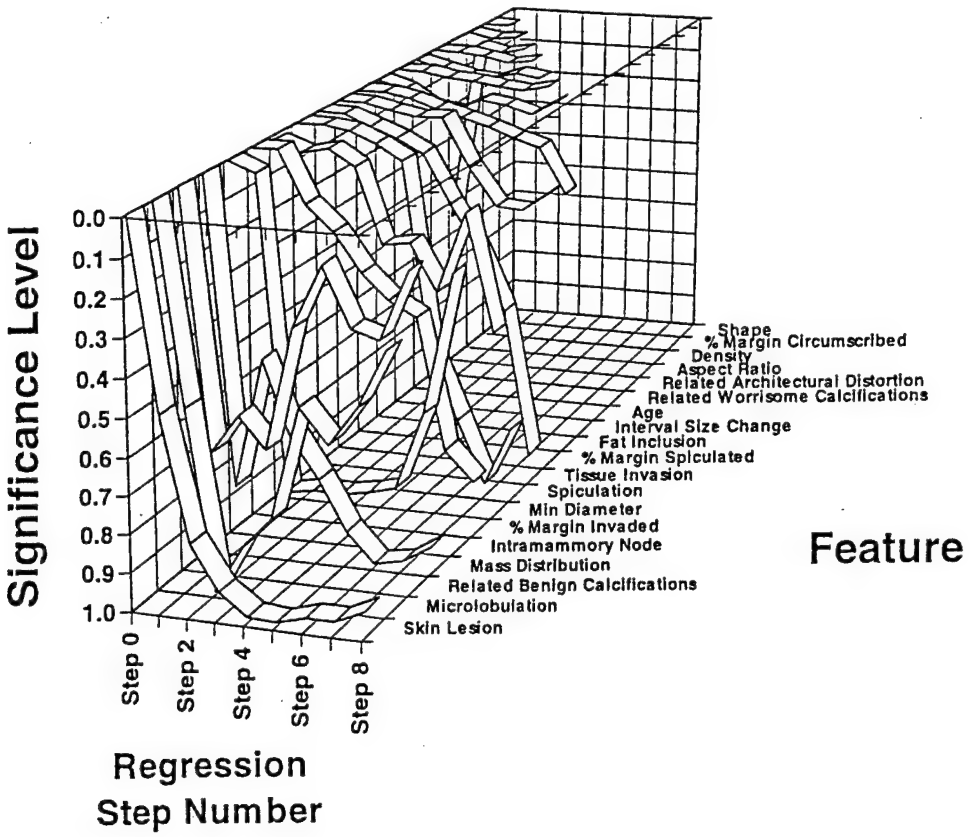


Figure 1. Stepwise logistic regression, BWH mass cases.

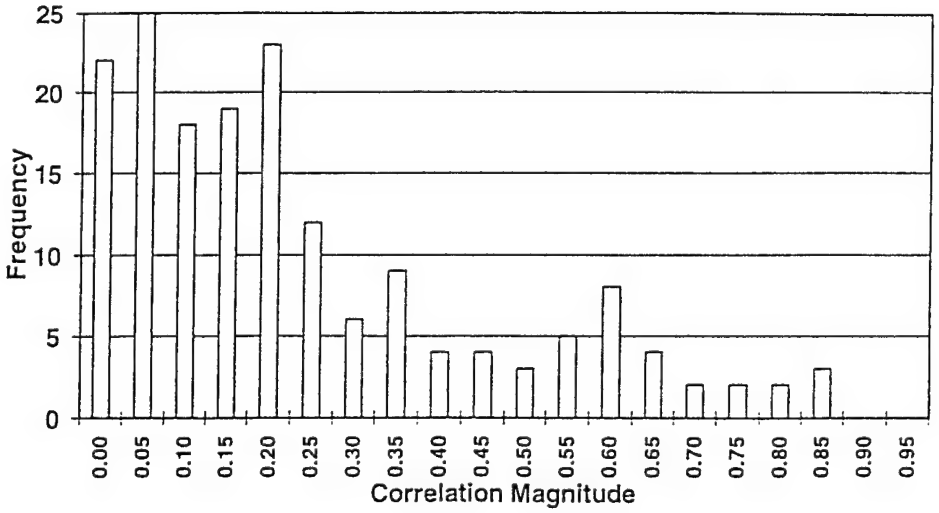


Figure 2. Intercorrelations of mass features, BWH cases.

The reason is apparent if we examine a histogram of the intercorrelations of pairs of features. Figure 2 shows the intercorrelations among mass features. The plot shows the number of feature pairs having different correlation values. Although many pairs of features have relatively low correlations, there are also many that have moderate to high correlations. Once a member of a highly correlated pair of features is introduced into the prediction rule, the additional diagnostic value of the other member becomes very small; one might say that the second feature is "old news."

The features selected for each SPR (masses and calcifications) are shown in Table IV – in the order of their ultimate significance. Features involving "change" were computed by the investigators from the readers' absolute judgments of the target and prior images. Note that two of the features (No. 6 for masses, No. 4 for calcifications) are what we have termed "interval-change" features. We will return to the question of the importance of such features.

Figure 3 shows the ROC for the SPR for masses that was based on BWH cases and readers. Figure 4, similarly, for calcifications. In these figures and in others showing ROCs both trained and tested (evaluated) on the same cases and readers, the SPRs are determined by a jackknifed, "leave-one-out" analysis, in which each single case/reader result being evaluated is eliminated from contributing to the SPR that is applied to that single result. This jackknife analysis is undertaken to reduce the optimistic bias that would be engendered by a common set of cases for training and test. The A_z value for the mass SPR, as shown in the inset of Fig. 3, is .94; the A_z for calcifications (Fig. 4) is .72. The 95% confidence bands for these A_z values are from .92 to .96 for masses and from .67 to .77 for calcifications. The superiority of a mass over a calcification SPR is consistent with findings of our earlier work; the features of calcifications are generally regarded as not being as reliably estimated as those of masses.

Table IV. Features Selected for BWH Statistical Prediction Rules

Rule		Features
For Masses		1. Shape of Mass
		2. Percent of margin circumscribed
		3. Size of Mass: computed ratio (maximum size/minimum size)
		4. Density of mass relative to surrounding glandular tissue
		5. Patient age
	*	6. Size of Mass: computed change (current study – prior study)
		7. Presence of related architectural distortion
		8. Presence of worrisome calcifications
For Calcifications		1. Presence of related architectural distortion
		2. Patient age
		3. Percent of tissue that is glandular
	*	4. Change in size of focal distribution over time: computed ((current study – prior study)/years between studies
		5. Degree to which the distribution can be characterized as segmental
		6. Degree to which the distribution can be characterized as linear
		7. Degree to which elements can be characterized as fine linear
		8. Degree to which elements can be characterized as pleomorphic
	*	An interval-change feature

The performances for masses and calcifications of the pooled BWH readers, and also the two corresponding SPRs, are given in Table V in terms of the ROC accuracy index A_z. The main result is given in column (1), which shows the accuracies of distinguishing malignant cases from all nonmalignant cases, i.e., both benign-biopsy cases and suspicious normal cases. Both readers and respective SPRs are in the low .90s for masses and in the low .70s for calcifications. The SPRs were no better than the readers. Having worked in a previous study (Getty, et al., 1988) just with cases confirmed by biopsy, where A_z values in the .80s were obtained, we analyzed such

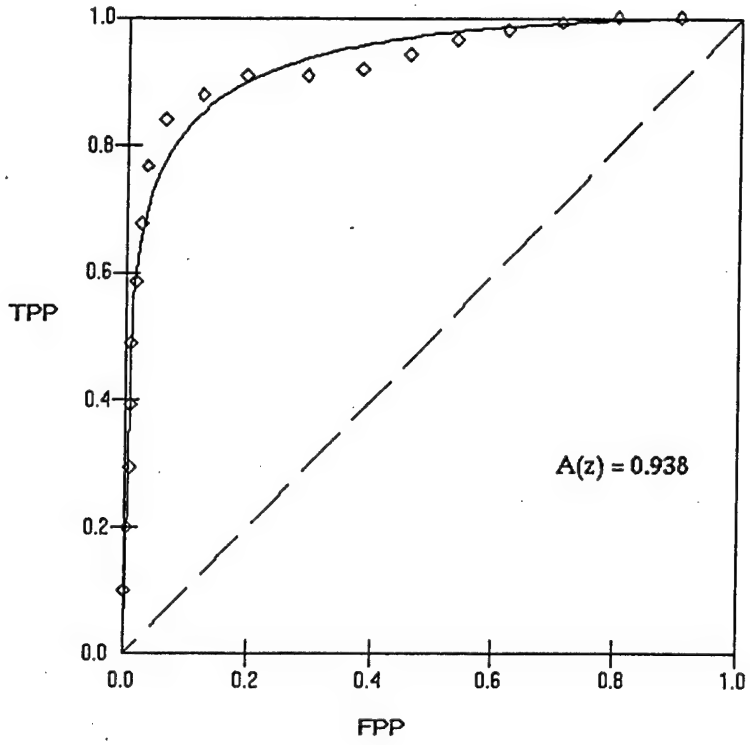


Figure 3. ROC for SPR for BWH mass cases (TPP = True-Positive Proportion, FPP = False-Positive Proportion).

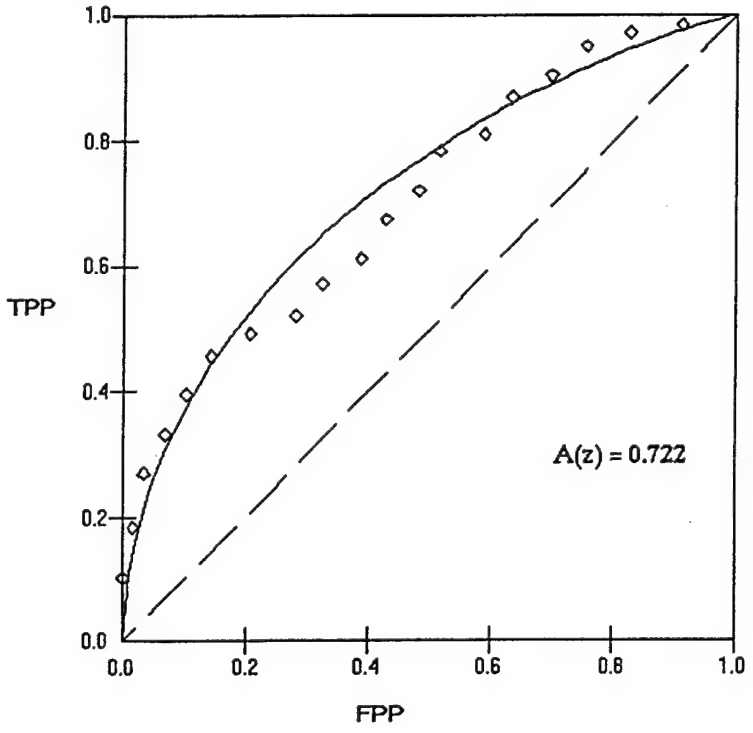


Figure 4. ROC for SPR for BWH calcification cases (TPP = True-Positive Proportion, FPP = False-Positive Proportion).

cases for masses, for which an adequate number of cases was available. As shown in column (2), we see again results in the .80s. Column (3) confirms that suspicious normal cases are easier than benign-biopsy cases.

Table V. Performances (A_z) of BWH Readers (Pooled) and Statistical Prediction Rules

		(1) Malignant vs. Benign-Biopsy and Suspicious Normal	(2) Malignant vs. Benign-Biopsy (Masses)	(3) Malignant vs. Suspicious Normal (Masses)
Masses (N ~ 510)	Readers	.94	.88	.96
	SPR	.94	.88	.97
Calcifications (N ~ 350)	Readers	.73		
	SPR	.72		

As an ancillary analysis, we measured the performances on the cases presenting as a mass for perceptual features taken from ultrasound imagery, a modality until recently used primarily to determine if masses are cysts and not to make the malignant/nonmalignant distinction. With recent advances in ultrasound technology and imagery, however, this modality is increasingly used to help differentiate malignancies and nonmalignancies. Table VI confirms with comparative figures that an SPR based only on ultrasound features performs very well (A_z = .93; 95% confidence limits of .91 and .95) as a diagnostic tool for the malignant/nonmalignant distinction. That figure differs from the SPR based on mammography features for the same subset of cases (A_z = .96; 95% confidence limits of .95 and .97), with $p<.05$.

Table VI. Performances of Ultrasound Features (A_z)

Readers	.94
SPR for cases with ultrasound features	.93
SPR for mammography features in same case set	.96
SPR for cases with both types of feature	.96

1.8 Results: HPHC baseline readings

Baseline (unaided) readings of 150 cases were made by five radiologists at HPHC, in anticipation of their later aided readings. The individual performances--in distinguishing malignant cases from both benign-biopsy and suspicious normal cases -- are given in column (1) of Table VII in terms of A_z. The mean A_z is .85. (Columns 2 and 3 were calculated for the purpose of feedback to the test mammographers: true-positive proportion at false-positive proportion = 0.5 and positive predictive value.)

Table VII. HPHC Baseline Reading Performance

Reader	(1) A_z	(2) TPP@FPP = .50	(3) PPV
1	.85	.90	.47
2	.91	.97	.49
3	.84	.89	.47
4	.82	.88	.47
5	.84	.93	.48
Mean	.85	.91	.48

With A_z based on pooled data (rather than the mean data of Table VII), the leftmost column of Table VIII gives a comparison of baseline HPHC performance on (1)

all cases, (2) masses alone, and (3) calcifications alone. (As usual, referring to Table VII, pooled data are seen to give slightly lower values of A_z than mean data.) The 95% confidence bands are .81 to .91 for masses and .73 to .85 for calcifications. The ROCs for masses and calcifications in the HPHC baseline condition are shown in Figs. 5 and 6, respectively.

Table VIII. A_z Values for HPHC Readers and SPR

	Baseline	Pooled Readers After Checklist	Pooled Readers After BWH SPR	HPHC SPR
All Cases	.83	.85	—	—
Masses	.86	.89	.89	.92
Calcifications	.79	.85	.85	.89

1.9 Results: enhanced reading study at HPHC

Preparations for conducting the enhanced reading study at HPHC included an initial briefing session for the mammographers as a group to go over the general procedure for the reading sessions. The written instructions prepared for the briefing are shown in Appendix B. This group session also included training on several of the calcification features that, based on the BWH results, we felt needed further explanation and illustration. The training images and feature data were taken from cases employed and readings generated in the BWH study.

The A_z values obtained in the HPHC enhanced readings—for the readers pooled and the SPRs—are shown in Table VIII. For all cases together, the readers'gain in A_z from "baseline" to "after checklist" was .02. For masses, this gain was .03 and, for calcifications, .06. The gain is significant only for calcifications, $p = .02$. The readers'

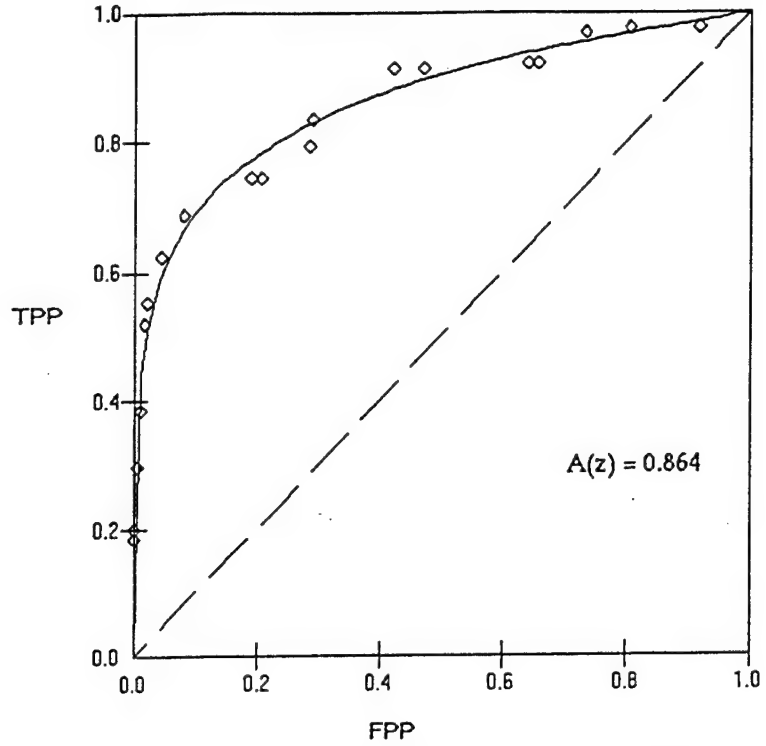


Figure 5. ROC for HPHC baseline reading of mass cases (TPP = True-Positive Proportion, FPP = False-Positive Proportion).

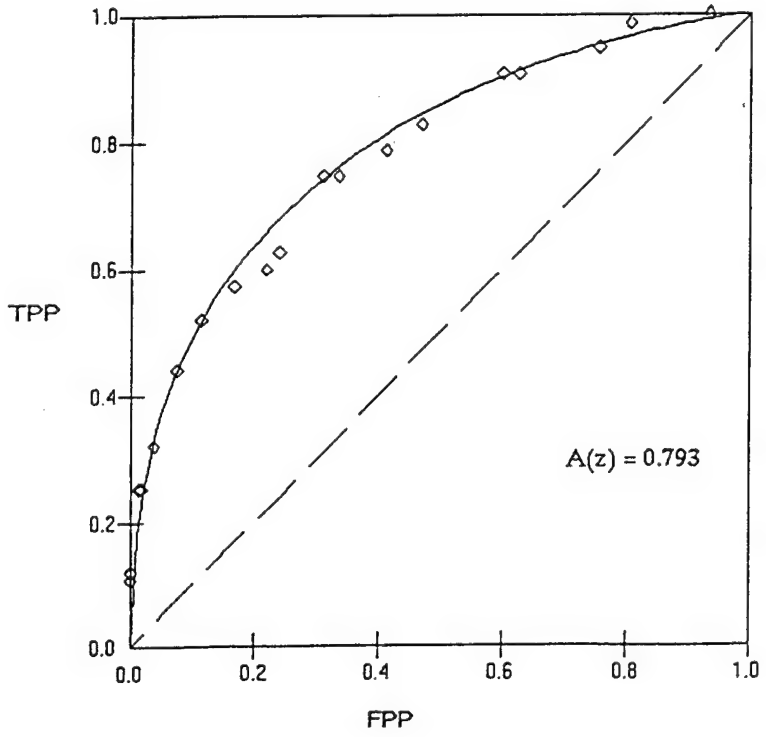


Figure 6. ROC for HPHC baseline reading of calcification cases (TPP = True-Positive Proportion, FPP = False-Positive Proportion).

performance after learning of the SPRs' estimates was not different from their performance when aided only by the checklist. The ROCs for the enhanced readings "after SPR" are shown for masses and calcifications in Figs. 7 and 8, respectively.

1.10 SPRs based on HPHC feature ratings

The feature ratings given in the enhanced readings at HPHC enabled calculating SPRs based on HPHC readers and cases. The A_z values for these SPRs (based on a jackknife analysis) are shown in the rightmost column of Table VIII -- .92 for masses and .89 for calcifications. Their ROCs are shown in Figs. 9 and 10.

The A_z for the SPR for masses based on HPHC ratings (.92) is .06 higher than baseline A_z for masses (.86); $p< .005$. For calcifications, the HPHC-based SPR has an A_z (.89) that is .10 higher the baseline value (.79); $p<.001$. The SPR accuracies are also significantly greater than the checklist-aided accuracies; $p<.001$ for masses and $p<.05$ for calcifications.

The features entering the HPHC-based SPRs are shown in Table IX, for masses and calcifications – in order of their ultimate significance. Again, features involving "change" were computed by the investigators from the readers' absolute judgments of the target and prior images. Here, six features are what we have termed "interval-change" features, substantially more than the two such features in the BWH-based SPRs as shown in Table III. For masses, the three most significant features are interval-change features; for calcifications, the most important feature is on this type.

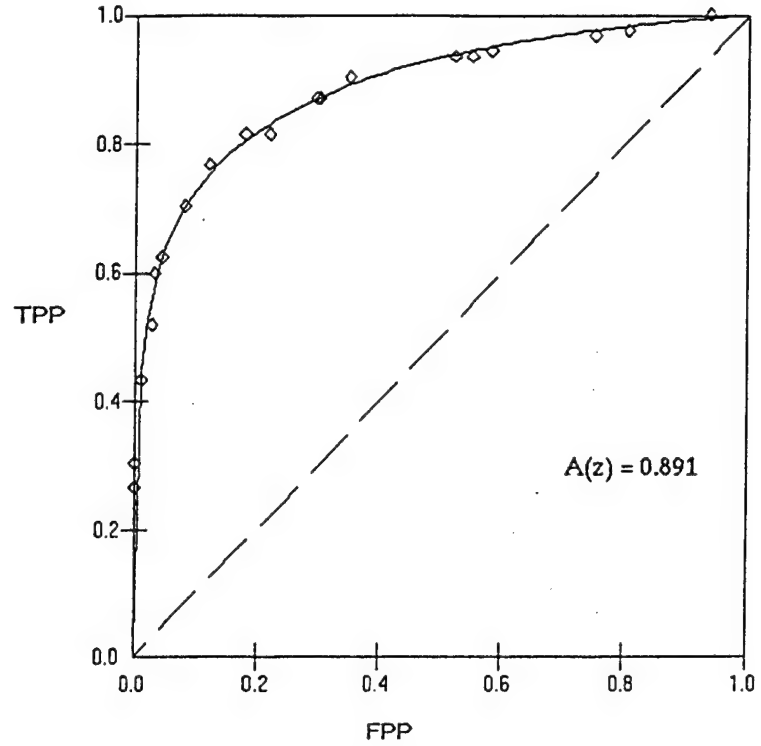


Figure 7. ROC for HPHC enhanced reading of mass cases, with BWH-based SPR (TPP = True-Positive Proportion, FPP = False-Positive Proportion).

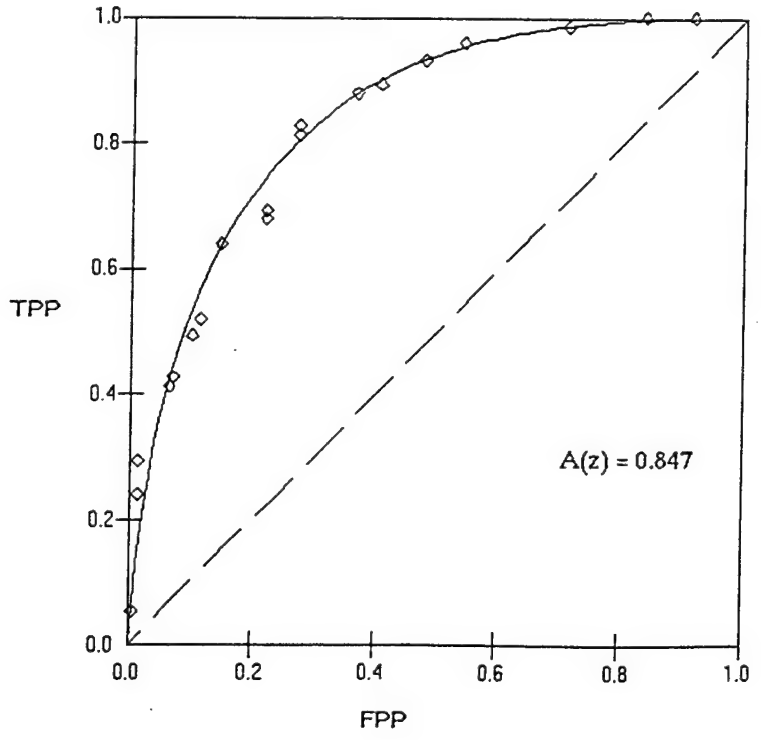


Figure 8. ROC for HPHC enhanced reading of calcification cases, with BHW-based SPR (TPP = True-Positive Proportion, FPP = False-Positive Proportion).

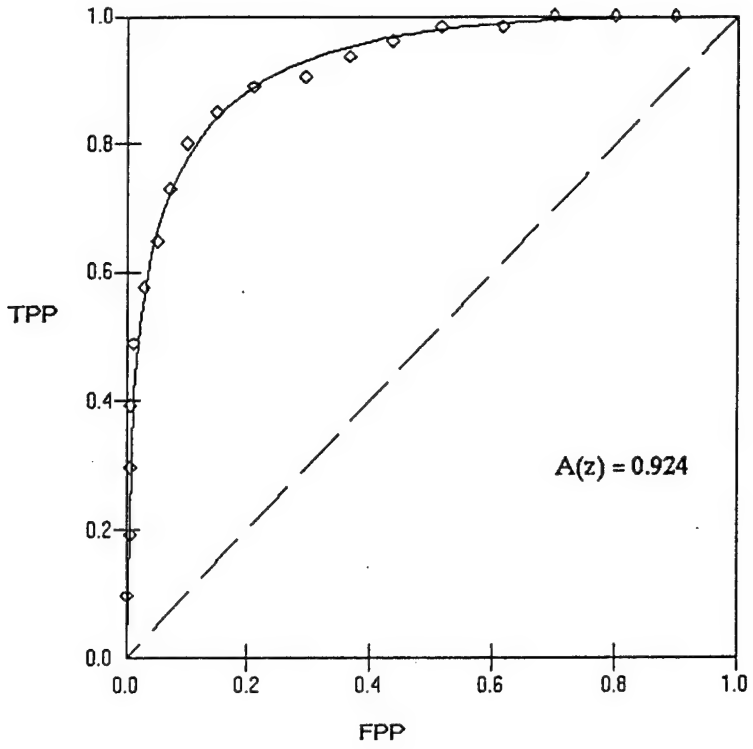


Figure 9. ROC for HPHC-based SPR for mass cases (TPP = True-Positive Proportion, FPP = False-Positive Proportion).

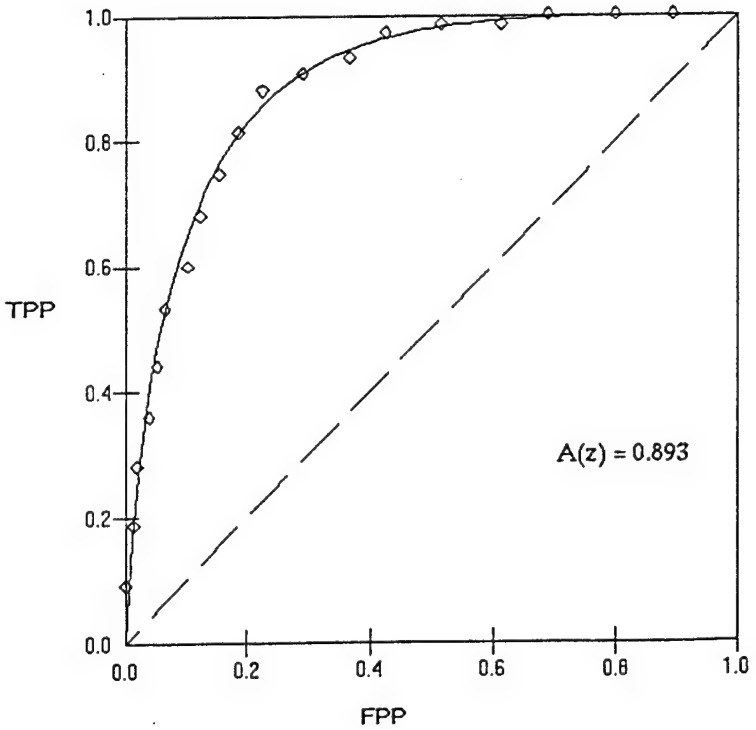


Figure 10. ROC for HPHC-based SPR for calcification cases (TPP = True-Positive Proportion, FPP = False-Positive Proportion).

Table IX. Features Selected for HPHC Statistical Prediction Rules

Rule		Features
For Masses	*	1. New; not significantly changed; significantly changed (three ordered categories)
	*	2. Rate of change in size (smallest diameter)
	*	3. Rate of change in aspect ratio (largest diameter/smallest diameter)
		4. Confidence regarding presence of related architectural distortion
		5. Largest diameter of mass (either view)
		6. Confidence that at least a small portion of border is spiculated
		7. Density of mass (relative to surrounding tissue)
	*	8. Rate of change in size of mass (largest diameter)
		9. Percentage of margin that is clearly circumscribed
		10. Patient age at imaging
For Calcifications	*	1. Amount of change in size (mm) of cluster in cranio-caudal view
		2. Degree to which elements characterized as pleomorphic (heterogeneous)
		3. Confidence regarding presence of related architectural distortion
	*	4. New; not significantly changed; significantly changed (three ordered categories)
		5. Number of elements in cluster
	*	An interval-change feature

1.11 Discussion

The development and testing of a decision-support system were accomplished essentially at the level of our scientific requirements and expectations. In one particular, the case sets ultimately obtained, though smaller than projected, were adequate. Our mammography consultants gave generously of their time and great capabilities. The checklist of features we developed was thorough and efficient. The SPRs contained reasonable numbers of significant features, both for masses and calcifications. Our varied statistical analyses were satisfactorily accomplished. Our

schedule of project deadlines prevented us from giving as much emphasis as we would have desired to interval-change features in the BWH-based SPRs, but allowed additional emphasis on them in the HPHC-based SPRs. The HPHC results for interval-change features fully justified the hypothesis -- heretofore based on anecdotal evidence -- that such features are of great importance in diagnosis and can be handled quantitatively in an SPR.

The enhancement value of the decision aids was less than expected: only the .06 increase for calcifications was statistically significant. The addition of the BWH-based SPR's advisory probability produced no further increase. A promising result is that the SPR for calcifications trained by HPHC feature ratings produced an accuracy of .89, thus showing an increase in accuracy of .10 over the HPHC baseline readings and of .17 over the BWH-trained SPR and, as well, a much higher absolute value for calcifications than obtained in any of our previous studies. This pattern of calcification results suggests that an iterative approach to SPR development would be fruitful. In such an approach, one would begin with the large, exhaustive set of candidate features (as we did with the BWH readers) and proceed to derive by statistical analysis the much smaller set used in the BWH-based SPR. This smaller set is more tractably discussed with readers (as we did with HPHC readers), who can then study those crucial features more effectively, and use them more consistently to generate a more accurate SPR and more accurate over-all performance.

2. Automated Report Writer

The function of the automated report writer is to take the feature ratings from the checklist/questionnaire and convert them into text that can substitute for the reports that would normally be dictated by the mammographer. One advantage of having the reports produced automatically in this fashion is that it can relieve the mammographer of the chore of dictating the report and compensate to a large extent for whatever extra effort is required in following and filling out the checklist/questionnaire. But the primary promise of the automated report lies in the potential for much more accurate and reliable communication of findings than can be achieved with typical dictated reports. This advance is possible because: 1) the automated reports will more reliably cover all of the essential diagnostic details, by drawing fully on the data collected on the checklist/questionnaire, and 2) those comprehensive findings will be reported in broadly recognized standard terms and phrasing (ACR, 1998).

Our overall effort in this area of the program of study was divided into two components: 1) development of the automated report writer and 2) evaluation of the reports with the assistance of a panel group of referring physicians (clinicians) and surgeons.

2.1 Development of the automated report writer

Although automated reports could easily be constructed just by chaining together a series of simple declarative sentences, each corresponding to an item on the checklist/questionnaire, such reports would tend to be clumsy to read and difficult to assimilate rapidly, and would be generally unacceptable to users in routine practice. The challenge, then, has been to design an automated process that can produce reports that approximate the flow of the mammographer's natural dictation. However, we also faced the critical need to accomplish that primary goal in as simple a way as possible.

Simplicity was critical not just to ensure that the work could fit into our limited budget of time and resources. We were also cautioned by our colleagues versed in computer-generated natural language to keep it simple as the most promising way to proceed: to provide the framework for adding levels of complexity, but to go as far as possible with the most primitive translational mechanisms first, adding complexity only as required to achieve adequate fluency (Meteer, 1991; 1992).

To keep our efforts further within practical bounds, we restricted development of the system to deal with the generation of reports of just one type of lesion, namely masses. Thus, all of the following discussion and various examples are couched in terms of features and descriptions of masses, and the evaluations were of reports concerned only with masses.

Because the database coming from the checklist/questionnaire is completely pre-specified (that is, no new fields or values occur while the system is in use), we used a simple "direct-replacement" grammar approach to generating the core parts -- sentences or clauses -- that comprise the report. Because all of the domain-specific information is encoded in the grammar, the same core system will generate reports of any kind, as long as the form of the input is similar. This arrangement allows the system to be extended to other types of mammography findings and potentially to other radiology contexts with few changes to the computer code. There are two advantages to this approach. First, all of the information specific to the domain, such as the order of the different parts of the report and the specific words to be used, are maintained in a declarative set of rules independent of the code that runs to produce the report. This independence makes easy small modifications to the report, such as changes in wording or additions of other information. Second, the system runs very fast, since at each point it only needs to select among a small number of alternatives,

and the computation required to make the selection is usually just a lookup in the database and comparison of numbers (e.g., the feature rating is less than four, or between five and seven, etc.).

We implemented this prototype system to generate reports off-line; that is, the information in the database is saved to an ASCII file that is input for the report writer, which produces all of the reports in batch mode. The system is written in LISP and runs on a Sparc Ultra, but it could be easily ported to a PC using a platform-independent language such as Java. We tested the system by running it on more than 500 cases, which the system processed in under 10 seconds per case.

We then conducted an informal, formative evaluation. We considered how adequately the system translated the various nuances of feature variation and how adequate the reports were with respect to their fluency and overall readability. Improving the capability of the system to convey nuances of feature variation was well accomplished by adjusting the thresholds that determined how the numerical data from the checklist/questionnaire is mapped onto the descriptive phrases available to the report writer for each feature. The more significant challenge was to determine how to improve fluency within and across sentences.

The major disadvantage of this simple replacement-grammar approach is the limit in the sophistication of the text one can generate. For example, because there is no link between different parts of the report or between different sentences, the wording is often repetitious. If, for example, the border of a mass shows evidence of both spiculation and invasion of adjacent tissue, two features that can be reported on independently in the checklist/questionnaire, they would, in the most primitive form of report, generate two independent sentences: 1) "The lesion shows evidence of spiculation" and 2) "The lesion shows evidence of tissue invasion." With very little

added complexity, but with a large gain in fluency, the process can be modified to combine sentences that would have the same stem. In this case: "The lesion shows evidence of spiculation and tissue invasion." Remarkably fluent reports of rather complex findings can be constructed by combining independently reported features (e.g., the size, shape, and locus of a lesion) as a chain of phrases within a single sentence. This approach is illustrated in the following example generated by the latest version of the report writer:

> Case #294 3/29/94
>
> The present examination is compared to a prior mammogram of 9/21/92. The breast is almost entirely fibroglandular.
>
> There is an 11 mm lobular mass at 2 o'clock in the posterior region of the left breast that has changed significantly since the last exam. It shows possible evidence of spiculation and likely evidence of tissue invasion. About half of the margin appears to be obscured by glandular tissue.
>
> An ultrasound was performed. The mass appears irregular with solid contents. The mass wall is irregular. The posterior wall of the mass displays shadowing.
>
> Impression: There is an 11 mm lobular mass in the left breast. This lesion is highly suggestive of malignancy. The probability of malignancy is .90, and appropriate action should be taken.

Clearly, such combined sentences can become clumsy if too many features are combined. However, some additional simple logic can correct for that. For example, the mechanism we initially built for generating the first sentence in the main body of the report tended to produce sentences that were a bit long, as exemplified in the second paragraph of the report shown above. Occasionally, it was challenged to include yet another qualifier, for example, the diagnostically important qualifier: "in a group of similar masses." The sentence then became too clumsy. This problem was solved by having the process test whether that occasional extra qualifier was present and, if so, to break the sentence in two.

In keeping with our initial commitment, we went as far as possible with the very simplest mechanisms and then added complexity, as exemplified above, to produce reports that were adequate in fluency to the exploratory aims of the present program.

2.2 Evaluation of the automated reports

Our general aim was to conduct a formative evaluation of the reports from the perspective of the physicians who will use the report – the clinician who requested the mammographic study and who will need to communicate the results to the patient, and the surgeon who will decide how to deal with a detected lesion. The general approach we followed was to obtain the two forms of report on each case -- the one as usually dictated and the other generated automatically from the checklist/questionnaire data.

2.2.1 Method. We first selected a representative set of eight mass cases and had our four mammographers read each of those cases on two independent occasions. On the first occasion, they read each case as they would normally and dictated the report as usual. On the second occasion, they read each case again, but this time following and filling out the checklist/questionnaire shown in Appendix C. (We separated the two activities to preclude an effect of the checklist/questionnaire on dictated reports, at the possible cost of not ensuring that the reader was similarly focused on the lesion in each case reading.) The questionnaire data on each case for each reader were then given to the automated report writer, which generated the reports. Materials and procedures were then prepared for evaluation of the reports by a panel of four physicians – two clinicians and two surgeons.

The evaluation was conducted during a two-hour panel meeting. We first asked each physician to evaluate the quality of each of 64 reports: the 2 forms of report on each of the 8 cases by each of the 4 mamographers. Specifically, the physicians gave a ten-point rating to each report on three different quantitative scales. Appendix C shows

one case as read by the four readers, along with the three scales. Scale *a* was for evaluating the degree to which the report was expressed in clear and standard terminology. Scale *b* was for evaluating adequacy of the amount of detail, whether just the right amount, too little, or too much. Scale *c* was for rating the fluency of the text. An additional rating form (also shown in Appendix C) was provided for each physician to rate consistency of the reports on each case across the four mammographers. This form provided for rating on two different scales: Scale *d* for rating consistency of the report with respect to terminology, and Scale *e* for rating consistency with respect to content.

After the physicians completed the quantitative ratings, which took approximately one hour, 45 minutes were devoted to open discussion. The physicians were encouraged to describe what they considered to be the critical functions of the report from their perspectives. They were also encouraged to critique the reports with respect to how well they met those critical needs and how they might be improved.

2.2.2 Panelist results: quantitative ratings. The results of the four physicians' ratings on five scales for both the dictated and automated reports are shown in Table X.

Table X. Mean Physician Ratings on Five 10-point Scales for Both Dictated and Automated Reports

Scale	Type of Report		Significance
	Dictated	Automated	
a. terminology	8.30	8.93	p<.001
b. detail	4.98	5.8	p<.001
c. fluency	8.08	6.47	p<.001
d. terminology	5.66	8.31	p<.001
e. content	6.08	6.02	p>.900

Analysis of variance showed the difference in mean ratings of the two types of report to be statistically significant at $p<.001$ for Scales *a, b, c, d* and the difference to be insignificant for Scale *e*. ($p>.90$).

For Scale *a* (clear and standard terminology), the automated report was deemed superior to the dictated report. The dictated report was rated better on Scale *b* (where the middle rating of 5 indicates the desired level of detail) and higher on Scale *c* (fluency). Across the four readings per case, the automated report rated higher on Scale *d* (consistency of terminology). The insignificant difference in Scale *e* was accompanied by a significant amount of case variation on that scale ($p<.001$). Case variation was otherwise insignificant, except for Scale *b* ($p<.02$).

Variation across readers on Scales *a, b, c* was insignificant (p's of about .80 to .90). Variation across physician panelists was significant for Scales *a, b, c* ($p<.001$) and insignificant for Scales *d, e* (p's of about .50 to .60). The difference in panelist occupation, between clinicians and surgeons, was insignificant for all five scales.

2.2.3 Panelist results: verbal critique. We began the critique by inviting the panelists to raise any points regarding strengths or weaknesses of the reports of either form that seemed most pressing to mention. This invitation elicited comments consistent across the panelists that showed their main concern was not with differences between the two forms of report, but rather with a common shortcoming of the conventional, dictated mammography report. The concern, one that we have previously encountered in our mammography studies, was that the mammography report commonly falls short in supporting the communications that have to be made to the patient and the actions that have to be taken in response to the report. The clinicians need support in conveying confidently to the patient what the report means, and both the clinicians and the surgeons need support with respect to deciding

confidently on the actions that they need to take. The details of the image-based lesion assessments that underlie the mammographer's diagnostic conclusion are not really wanted by the clinician or surgeon users. From the clinician/surgeon perspective, those data are mainly for the mammographer to digest into the diagnosis, not to hand on to them. Any more detail in the report beyond the very fewest essentials that support the diagnosis are seen as potentially confusing and often as a basis for hedging the diagnosis. While this criticism is understandable, in the sense that any report short of plain and simple is likely to be difficult to interpret and deal with, it is basically unrealistic. The mammography report surely should be kept as simple and definitive as possible, but if the situation is indeed complex and cannot accurately be reduced to a simple and definitive report, then all in the chain of communication, including the patient, have to be prepared to deal with it. While we did not attempt to debate or resolve the issue, we did consider it a theme important to have aired and potentially useful in interpreting the panelists' subsequent comments on the strengths and weaknesses of the two forms of report. We then proceeded to invite those comments.

The panelists were generally not prepared to come down on the side of one form or the other. Consistent with the quantitative data, the automated report was considered to be expressed in more objective terms, contributing to a sense of greater consistency across the reports. Also, the automated report was considered to be more stilted than the dictated report, again consistent with the quantitative finding. Perhaps the most informative distinction between the two forms of report was in regard to how well they showed the connection between the described findings and the diagnostic impression. The dictated reports were favored for showing a clearer connection. The automated reports had two strikes against them in this regard. First, the descriptive section of the automated reports impressed the panelists as having a lot more detail

than in the dictated reports, and that in itself would be seen as a drawback in light of their general concern that reports should be kept as simple as possible. Second, the automated impressions, while very succinct and to the point, an aspect presumably favorable in light of the panelists general yearning for simplicity, was seen as too skeletal, with insufficient connection to the preceding elaborate description.

The general point seems to be -- keep it simple, but when you give details, their connection to the impression should be obvious or explicit. Connectivity then is a very important consideration in developing the automated report. If we are prepared to argue that the mammographer should not be required to sacrifice accuracy for simplicity, we must also be prepared to argue that the mammographer make sure that the connections between the descriptive details and the diagnostic impression be abundantly clear. The mammographer is apparently better at that in conventional dictation; our automated process obviously needs improvement in this regard. Another concern that the panelists had with the automated reports is that the impressions all end with a statement of the probability of the diagnosis. Perhaps the clinician/surgeon sees giving the probabilistic expression of the diagnosis as a form of hedging. But once again, it can be argued that the mammographer should not be required to sacrifice accuracy for simplicity -- in this case turning a correctly stated shade of gray into an intolerably risky statement in black or white. The solution may be to provide for a better way of expressing the uncertainty. As one of the panelists put it, the expression of a probability however small, that something is bad, instills substantial concern in the patient. Perhaps a better language for expressing the uncertainty in degrees, but in less threatening qualitative terms, should be sought. One advantage of the automated report is that it might enable each end user to choose among alternative mappings of the diagnostic probability issued by the mammographer onto a continuum

of verbal renditions designed to treat patient concern appropriately in the particular setting.

2.3 Discussion

Development and evaluation of the report writer were accomplished as planned, and the several advances in understanding that we had sought in this problem area were achieved. The automated reports were found to be superior to the dictated reports in several aspects related to the standardization and precision of descriptive terms and to the consistency of usage across mammographers. Though the automated reports were judged to be a little more stilted than the dictated ones, we found that the automated reports were remarkably good in regard to the complexity of expression that they could smoothly support as well as having good overall fluency.

As to the panelists' general concern that mammography reports should be simpler and more definitive, we have a much less clear sense of insight and of how to respond. Clearly, the mammographer can not be expected to sacrifice accuracy, in the form of needed detail of description and needed shading of interpretation, just to make the clinician, surgeon, or patient more comfortable in dealing with the report. On the other hand, keeping the report as simple as possible, expressing the shading of interpretation in degrees no finer than necessary, and keeping all expression tightly mapped to a standard vocabulary and format can help to make the report easier to read and interpret. Automated report writing should help in this regard. Thus, even this more general concern looks to be more promising to address in light of our present success with the automated report.

CONCLUSIONS AND FUTURE DEVELOPMENTS

The clinical evaluation of a decision-support system for mammogram interpretation that was made in this project yielded less of an over-all increase in reading accuracy than observed in previous laboratory studies. One should keep in mind, however, though not specifically shown in the present study, that such a system has been found to yield larger increases in accuracy for the more difficult cases (Swets, et al., 1991). Moreover, several results of this evaluation give promise of practical import and serve to recommend further research and development. One gratifying result is the demonstration that accuracy for cases presenting as calcifications can be increased to about the level observed before only for cases presenting as masses.

Another outcome of note is the indication that the gain in accuracy of the system came largely from its checklist component. The checklist, of course, can be used in practice at minimal expense and complexity, without the computer-based merging of checklist values into a diagnostic probability. Meanwhile, even if the SPR lends little or no further accuracy increase, it nonetheless serves to present a "second opinion" in precise quantitative terms, if only to compare with the reader's own estimate of the probability of malignancy.

When the checklist is used and computer support is available, the automated report writer can be used to improve reporting consistency within and across radiologists and to ensure a desired format and vocabulary, in particular, the vocabulary recommended by the American College of Radiologists.

A next step would be to make the decision-support system and automated report writer available freely on the Internet. Wide usage would make more accurate the estimates of statistical prediction rules, by virtue of very large sample sizes on which to train them, and could provide SPRs tailored to the case mixes of specific sites. Differing

opinions between readers even at different sites could be discussed in the uniform, quantitative terms of well-defined perceptual features. Ongoing tutorials could be tailored to the varying, specific needs of different radiologists as evidenced by their responses to chosen teaching cases (Greenes, Swets, Getty, and Pickett, 1992). Reports of case findings could be distributed to referring physicians and surgeons as they are developed. These report users would have a clear focus in the uniformly formatted reports for a collaborative effort to improve reporting for their needs.

In general, we propose that the explicit treatment of diagnosis and reporting that is provided in the system described and evaluated in the project will lead to increased quality assurance. It would naturally be incorporated into a larger medical information system that would also include decision support for test ordering; approval mechanisms; test scheduling; and workflow management; as well as connections to patient records.

To be sure, the decision-support and reporting system would have to be convenient to use. Radiologists who read the minimum number of mammograms required to maintain certification, about 10 per week, may not desire to adopt any special system. Radiologists who read 100 per week (as do our HPHC radiologists) will be very concerned not to increase the time per reading. For them, asking a colleague for a second opinion may be the preferred way to handle the relatively infrequent difficult case. The usefulness of the proposed system may be largely confined to the radiologists inbetween.

Further, the increase in accuracy provided by the system would have to be appreciated by radiologists, particularly if they are operating at the upper right of the ROC, where an increase in accuracy is likely to translate into a reduction of false-positive decisions regarding follow-up imaging or biopsy. The paradoxical possibility exists in mammographic interpretation that a false-positive decision is good rather than bad: the

newer needle biopsies are relatively innocuous and more frequent and patients with negative biopsies are found actually to be appreciative, rather than offended by an "unnecessary" biopsy. Of course, patients in whom a small tumor is found early tend to be most appreciative, so the true-positive decision is still most highly valued. The low yield of biopsies – that is, positive predictive value in the middle .20s (J. E. Meyer, BWH, personal communication, October, 1998) – is indicative of the lenient decision threshold used to go on to biopsy. This threshold is not necessarily cavalier with respect to spending the nation's health dollars; two return visits for follow-up imagery may cost as much. And the delay till a good outcome may serve to inflict as much patient concern.

If for such reasons, the proposed system is not widely adopted, we would point up its potential value as a teaching tool. The radiologists on this project are appreciative of the lessons learned from our statistical analysis of perceptual features, i.e., which features are diagnostic to what extent, which ones are relatively independent and which are highly correlated, and how to rate them quantitatively. Of special note are this project's scientific confirmation of interval-change features and demonstrated refinement of calcification features. At a time when the American College of Radiology and the American Cancer Society are working together to improve consistency of mammographic interpretations within and between radiologists, the systematic approach of a statistical analysis of perceptual features may well turn out to be the basis of the most effective intervention. We have given previous feature results to an ACR/ACS team and will make our new results available to it. The use of our perceptual features in training may pave the way for their use in practice.

REFERENCES AND BACKGROUND LITERATURE

ACR (1998) Illustrated Breast Imaging Reporting and Data System (BI-RADS), 3RD edition.

D'Orsi, C. J., Getty, D. J., Swets, J. A., Pickett, R. M., Seltzer, S. E., and McNeil, B. J. (1992) Reading and decision aids for improved accuracy and standardization of mammographic diagnosis. Radiology, 184, 619-622.

Dorfman, D. D., Berbaum, K. S., and Metz, C. E. (1992) Receiver operating characteristic rating analysis: Generalization to the population of readers and patients with the jackknife method. Investigative Radiology, 27, 723-731.

Getty, D. J., Pickett, R. M., Chylack, L. T., McCarthy, D. F., and Huggins, A.W.F. (1989) An enriched set of features of nuclear cataract identified by multidimensional scaling. Current Eye Research, 8(1), 1-8.

Getty, D. J., Pickett, R. M., D'Orsi, C. J., and Swets, J. A. (1988) Enhanced interpretation of diagnostic images. Investigative Radiology, 23(4), 240-252.

Greenes, R. F., Swets, J. A., Getty, D. J., Pickett, R. M. (1992) Computer assisted instruction in radiology. Rinal Report submitted to Brigham and Women's Hospital. National Institute of Health, Grant No. 5 R01 CA45574-03.

Hosmer, D. W. and Lemeshow, S. (1989) Applied logistic regression. New York: Wiley.

Kopans, D. and D'Orsi, C. J. (1992) ACR system enhances mammography reporting. Diagnostic Imaging, September, 125-132.

Meteer, M. (1991) Bridging the "generation gap" between text planning and linguistic realization. Computational Intelligence, 7(4).

Meteer, M. (1992) Expressibility and the problem of efficient text planning. London: Pinter.

Seltzer, S.E., McNeil, B. J., D'Orsi, C. J., Getty, D. J., Pickett, R. M., and Swets, J. A. (1992) Combining evidence from multiple imaging modalities: A feature-analysis method. Computerized Medical Imaging and Graphics, 16(6), 373-380.

Swets, J. A. (1979) ROC analysis applied to the evaluation of medical imaging techniques. Investigative Radiology, 14(2), 109-121.

Swets, J. A. (1986a) Indices of discrimination or diagnostic accuracy: Their ROCs and implied models. Psychological Bulletin, 99(1), 100-117.

Swets, J. A. (1986b) Form of empirical ROCs in discrimination and diagnostic tasks: Implications for theory and measurement of performance. Psychological Bulletin, 99(2), 181-198.

Swets, J. A. (1988) Measuring the accuracy of diagnostic systems. Science, 240, 1285-1293.

Swets, J. A. (1992) The science of choosing the right decision threshold in high-stakes diagnostics. American Psychologist, 47(4), 522-532.

Swets, J. A. (1996) Signal detection theory and ROC analysis in psychology and diagnostics. Mahwah, N J: Erlbaum.

Swets, J. A., Getty, D. J., Pickett, R. M., D'Orsi, C. J., Seltzer, S. E., and McNeil, B. J. (1991) Enhancing and evaluating diagnostic accuracy. Medical Decision Making, 11, 9-18.

Swets, J. A. and Pickett, R. M. (1982) Evaluation of diagnostic systems: Methods from signal detection theory. NY: Academic Press.

Swets, J. A., Pickett, R. M., Whitehead, S. F., Getty, D. J., Schnur, J. A., Swets, J. B., Freeman, B. A. (1979) Assessment of diagnostic technologies. Science, 205, 753-759.

Appendix A: Five Modules of Checklist

Reader No._____

Case No._____

Finding No._____

Response Form--X-Ray Mammography

Overview of Breast Images

- Percentage of Tissue that is Glandular

  Current _____ % OV02

- Identify the finding FNDG1

  - O Mass
  - O Not-definitely-benign calcifications
  - O Asymmetric Breast Tissue
  - O Architectural Distortion
  - O Regional Calcifications

Module I

Reader No._____

Case No._____

Finding No._____

Mass (MM)

Relationship to Prior Study

- This mass finding is: MM20

  - O new
  - O not significantly changed
  - O significantly changed

If significantly changed, also rate the prior images where requested. Otherwise, rate only the current images.

- Density of mass relative to surrounding glandular tissue

Current 0 1 2 3 4 5 6 7 8 9 10 MM04

mass density much lower — isodense — mass density much higher

- Confidence about the presence of fat within the mass

Current 0 1 2 3 4 5 6 7 8 9 10 MM05

definitely NONE present — definitely some present

- Size of mass

Current :	Largest diameter (in either CC or oblique view)	_____ mm	MM06
Current :	Smallest diameter (in either CC or oblique view)	_____ mm	MM07
Prior:	Largest diameter (in either CC or oblique view)	_____ mm	MM06
Prior:	Smallest diameter (in either CC or oblique view)	_____ mm	MM07

Mass (MM) - cont.

- Shape of mass

Current 0 1 2 3 4 5 6 7 8 9 10 MM10
round/oval lobular irregular

Prior 0 1 2 3 4 5 6 7 8 9 10

- Percentage of the margin that is clearly circumscribed

Current _____ % MM13A

Prior _____ %

- Confidence that at least a small portion of the margin is spiculated

Current 0 1 2 3 4 5 6 7 8 9 10 MM12
definitely NOT spiculated — definitely spiculated

- Confidence that the mass is an intramammary node

Current 0 1 2 3 4 5 6 7 8 9 10 MM16
definitely NOT an intramammary node — definitely an intramammary node

- Confidence regarding the presence of related architectural distortion

Current 0 1 2 3 4 5 6 7 8 9 10 MM17
definitely NOT present — definitely present

- Confidence regarding the presence of worrisome calcifications within the mass

Current 0 1 2 3 4 5 6 7 8 9 10 MM18
definitely NOT present — definitely present

Mass (MM) - cont.

Initial Overall Diagnostic Judgment

- **Benign vs. Malignant**
  Rate the likelihood (as the number of chances in 100) that the finding is indicative of malignancy:

  Rating (0 to 100) _____ **MMRA1**

  where: 0 = certainly benign or normal

  100 = certainly malignant

Computed Probability of Malignancy: _____

Final Overall Diagnostic Judgment

- **Benign vs. Malignant**
  Rate the likelihood (as the number of chances in 100) that the finding is indicative of malignancy:

  Rating (0 to 100) _____ **MMRA2**

  where: 0 = certainly benign or normal

  100 = certainly malignant

Module II

Reader No._____

Case No._____

Finding No._____

Calcifications (Not-Definitely-Benign) (NC)

Relationship to Prior Study

- This not-definitely-benign calcifications finding is: NC21

  O new

  O not significantly changed

  O significantly changed

If significantly changed, also rate prior images where requested. Otherwise, rate only the current images.

Element Characteristics

- Size of largest individual element (best visual estimate)

Current	O	O	O	NC05
	less than 0.5 mm	0.5 mm to 1.0 mm	more than 1.0 mm	

- Variability of size of elements

Current	0	1	2	3	4	5	6	7	8	9	10	NC06
	low variability of size									high variability of size		

Calcifications (Not-Definitely-Benign)(NC) - cont.

- Degree to which the elements can be characterized as fine linear

Current 0 1 2 3 4 5 6 7 8 9 10 NC07

definitely NONE of the elements are fine linear

at least one or two elements definitely are, or several probably are, fine linear

Prior 0 1 2 3 4 5 6 7 8 9 10

- Degree to which the elements can be characterized as branching

Current 0 1 2 3 4 5 6 7 8 9 10 NC08

definitely NONE of the elements are branching

at least one or two elements definitely are, or several probably are, branching

Prior 0 1 2 3 4 5 6 7 8 9 10

- Degree to which the elements can be characterized as pleomorphic (heterogeneous)

Current 0 1 2 3 4 5 6 7 8 9 10 NC09

definitely NONE of the elements are pleomorphic (heterogeneous)

at least one or two elements definitely are, or several probably are, pleomorphic (heterogeneous)

Prior 0 1 2 3 4 5 6 7 8 9 10 • Degree the elements can be characterized as punctate

Current 0 1 2 3 4 5 6 7 8 9 10 NC11 definitely

definitely NONE of the elements are punctate

at least one or two elements definitely are, or several probably are, punctate

Distribution Characteristics

- Number of elements

Current ○ less than 5 ○ 5 to 10 ○ more than 10 NC13

Calcifications (Not-Definitely-Benign)(NC) - cont.

- Size of the focal distribution

  Largest dimension in CC view

  Current _____ mm NC14

  Prior _____ mm NC14

- Degree to which the distribution can be characterized as linear

Current	0	1	2	3	4	5	6	7	8	9	10	NC17
	definitely NOT linear										definitely linear	
Prior	0	1	2	3	4	5	6	7	8	9	10	

- Degree to which the distribution can be characterized as segmental

Current	0	1	2	3	4	5	6	7	8	9	10	NC18
	definitely NOT segmental										definitely segmental	
Prior	0	1	2	3	4	5	6	7	8	9	10	

Relationship to Other Aspects of This Study

- Confidence regarding presence of related architectural distortion

Current	0	1	2	3	4	5	6	7	8	9	10	NC19
	definitely NOT present										definitely present	

- Confidence regarding presence of related mass or asymmetric breast tissue

Current	0	1	2	3	4	5	6	7	8	9	10	NC20
	definitely NOT present										definitely present	

Calcifications (Not-Definitely-Benign)(NC) - cont.

Initial Overall Diagnostic Judgment

- **Benign vs. Malignant**
  Rate the likelihood (as the number of chances in 100) that the finding is indicative of malignancy:

  Rating (0 to 100) _____ **NCRA1**

  where: 0 = certainly benign or normal

  100 = certainly malignant

Computed Probability of Malignancy: _____

Final Overall Diagnostic Judgment

- **Benign vs. Malignant**
  Rate the likelihood (as the number of chances in 100) that the finding is indicative of malignancy:

  Rating (0 to 100) _____ **NCRA2**

  where: 0 = certainly benign or normal

  100 = certainly malignant

Module III

Reader No._____

Case No._____

Finding No._____

Asymmetric Breast Tissue (AT)

Relationship to Prior Study

- This asymmetric tissue finding is: AT10

  - O new
  - O not significantly changed
  - O significantly changed

If significantly changed, also rate the prior images where requested. Otherwise, rate only the current images.

- Size of distribution of asymmetric breast tissue

	Current	Prior	
Largest diameter in CC view	_____ mm	_____	AT05
Largest diameter in oblique view	_____ mm	_____	AT06

- Confidence regarding the presence of worrisome calcifications within the asymmetric breast tissue

Current 0 1 2 3 4 5 6 7 8 9 10 AT07
definitely NOT present — definitely present

Prior 0 1 2 3 4 5 6 7 8 9 10

- Confidence regarding presence of related architectural distortion

Current 0 1 2 3 4 5 6 7 8 9 10 AT09
definitely NOT present — definitely present

Prior 0 1 2 3 4 5 6 7 8 9 10

Asymmetric Breast Tissue (AT) - cont.

Overall Diagnostic Judgment

- **Benign vs. Malignant**
  Rate the likelihood (as the number of chances in 100) that the finding is indicative of malignancy:

  Rating (0 to 100) _____ **ATRA**

  where: 0 = certainly benign or normal

  100 = certainly malignant

Module IV

Reader No._____

Case No._____

Finding No._____

Architectural Distortion (AD)

Relationship to Prior Study

- This architectural distortion finding is: AD11

  - O new
  - O not significantly changed
  - O significantly changed

If significantly changed, also rate the prior images where requested. Otherwise, rate only the current images.

- Confidence that the architectural distortion is related to prior surgery

Current	0	1	2	3	4	5	6	7	8	9	10	AD06
	definitely NOT related to prior surgery									definitely related to prior surgery		
Prior	0	1	2	3	4	5	6	7	8	9	10	

- Confidence regarding the presence of related worrisome calcifications

Current	0	1	2	3	4	5	6	7	8	9	10	AD07
	definitely NOT present									definitely present		
Prior	0	1	2	3	4	5	6	7	8	9	10	

Architectural Distortion (AD) - cont.

- Confidence regarding the presence of a related mass

Current	0	1	2	3	4	5	6	7	8	9	10	AD09
	definitely NOT present									definitely present		
Prior	0	1	2	3	4	5	6	7	8	9	10	

- Confidence regarding the presence of related asymmetric breast tissue

Current	0	1	2	3	4	5	6	7	8	9	10	AD10
	definitely NOT present									definitely present		
Prior	0	1	2	3	4	5	6	7	8	9	10	

Overall Diagnostic Judgment

- **Benign vs. Malignant**

  Rate the likelihood (as the number of chances in 100) that the finding is indicative of malignancy:

  Rating (0 to 100) _____ **ADRA**

  where: 0 = certainly benign or normal

  100 = certainly malignant

Module V

Reader No._____

Case No._____

Finding No._____

Regional Calcifications (Not-Definitely-Benign) (RC)

Relationship to Prior Study

- This regional calcifications finding is: RC21

  - O new
  - O not significantly changed
  - O significantly changed

If significantly changed, also rate the prior images where requested. Otherwise, rate only the current images.

Element Characteristics

- Size of largest individual element (best visual estimate)

  Current O less than 0.5 mm O 0.5 mm to 1.0 mm O more than 1.0 mm RC05

- Variability of size of elements

  Current 0 1 2 3 4 5 6 7 8 9 10 RC06

  0 = low variability of size; 10 = high variability of size

- Degree to which the elements can be characterized as fine linear

  Current 0 1 2 3 4 5 6 7 8 9 10 RC07

  0 = definitely NONE of the elements are fine linear; 9 or 10 = at least one or two elements definitely are, or several probably are, fine linear

  Prior 0 1 2 3 4 5 6 7 8 9 10

Calcifications (Not-Definitely-Benign) (RC) - cont.

- Degree to which the elements can be characterized as branching

Current 0 1 2 3 4 5 6 7 8 9 10 RC08

definitely NONE of the elements are branching

at least one or two elements definitely are, or several probably are, branching

Prior 0 1 2 3 4 5 6 7 8 9 10

- Degree to which the elements can be characterized as pleomorphic (heterogeneous)

Current 0 1 2 3 4 5 6 7 8 9 10 RC09

definitely NONE of the elements are pleomorphic (heterogeneous)

at least one or two elements definitely are, or several probably are pleomorphic (heterogeneous)

Prior 0 1 2 3 4 5 6 7 8 9 10 ●

to which the elements can be characterized as punctate

Current 0 1 2 3 4 5 6 7 8 9 10 RC11

NONE of the

at least one or two elements are punctate

elements definitely are, or several probably are punctate

Relationship to Other Aspects of This Study

- Confidence regarding presence of related architectural distortion

Current 0 1 2 3 4 5 6 7 8 9 10 RC19

definitely NOT present

definitely present

- Confidence regarding presence of related mass or asymmetric breast tissue

Current 0 1 2 3 4 5 6 7 8 9 10 RC20

definitely NOT present

definitely present

Regional Calcifications (Not-Definitely-Benign) (RC) - cont.

Overall Diagnostic Judgment

- **Benign vs. Malignant**

Rate the likelihood (as the number of chances in 100) that the finding is indicative of malignancy:

Rating (0 to 100) _____

where: 0 = certainly benign or normal

100 = certainly malignant

Appendix B

Mammography Research Program
HPHC Reading Aid Study (January - February 1998)

Instructions

This research program is concerned with testing a system for enhanced mammographic diagnosis of breast cancer. The enhancement consists of two parts. The first is a checklist/questionnaire that reminds the mammographer of all of the important diagnostic features of a worrisome lesion to evaluate and that provides for an explicit report on the type, status or extent of each feature. The second part is a statistical prediction rule that takes as input the various feature reports on the questionnaire and merges them optimally into an advisory report back to the mammographer on the probability that the lesion is malignant.

Your role in the program is to help in testing the effectiveness of this system. You have already read 150 cases following your standard reading procedure. Now we want to have you reread those cases following the enhanced procedure so that we can determine, by comparing your performance under the two conditions, how well it works.

We expect that just working with a checklist and having to consider explicitly, and rate quantitatively, each of the important diagnostic features on each case will be beneficial in itself. We expect, however, that the primary benefit will come from your bottom line consideration of the advisory probability of malignancy provided on each case by the statistical prediction rule. But the critical requirement for gaining that benefit is for you to provide the prediction rule with an assessment of each feature that allows it to speak as strongly as it can for itself. Thus, it is very

important that your judgments of each feature be made as carefully and independently as possible. This is particularly important if a principal sign is present that tilts you strongly toward a malignant decision. Decisions based on obvious signs can be wrong, and the information gathered on the other features may help to protect against such errors. If you leap to a wrong overall judgment and as a result take less care in reading the other features, you deprive them of the opportunity to protect you from making that error.

Also, in regard to several features that require judgments of both current and prior status, it may be tempting, when there looks to be hardly any difference, to make a careful measurement on one, and then to call the other the same. In so doing, you are throwing away some information in that second rating, information that may be informative. Similarly, in regard to the separate judgments we ask you to make about a mass regarding the presence of spiculation and the presence of an indistinct border. Try to avoid letting a presumption that the two go together reduce the care and independence of those separate judgments. The strength of the feature-based approach to enhancement lies in getting as much independent information as possible out of each feature.

The questionnaire has been designed to support detailed reports on worrisome lesions in all of their main radiographic presentations, including as a mass, as architectural distortion, as calcifications, and as an asymmetric density. To simplify its application, we have developed separate modules that cover just those features relevant to each presentation. These modules can be used individually if a lesion presents, say, just as a mass, or just as calcifications, or in combination if, say, it presents as a mass with calcifications. For each case we will provide you with the questionnaire module(s) relevant to its particular radiographic presentation(s). Your task will be to fillout those questionnaires as carefully as possible.

We will be pleased to answer any questions you have about the overall design of the questionnaires, the wording and structure of particular feature questions or about this study or the general program. Details of the actual reading procedure will be explained at your first reading session.

Appendix C
Questionnaire Used to Generate Automated Reports

Reader _____

Case No._____

Response Form--X-Ray Mammography
(Condition 2 of Report Writer Study)

Overview of Breast Images

- Percentage of Tissue that is Glandular

  Current _____ % OV02

Mass Findings

- This mass finding is: MM20

  O new

  O not significantly changed

  O significantly changed

- Distribution

  Current

  O single mass

  O multiple similar masses MM02

- Density of mass relative to surrounding glandular tissue

Current 0 1 2 3 4 5 6 7 8 9 10 MM04

mass density much lower — isodense — mass density much higher

- Size of mass

	Current	Prior	
Largest diameter in CC view	_____ mm	_____	MM06
Smallest diameter in CC view	_____ mm	_____	MM07
Largest diameter in oblique view	_____ mm	_____	MM08
Smallest diameter in oblique view	_____ mm	_____	MM09

- Shape of mass

Current 0 1 2 3 4 5 6 7 8 9 10 MM10

round/oval — lobular — irregular

- Confidence that at least a small portion of the margin is indistinct due to tissue invasion

Current 0 1 2 3 4 5 6 7 8 9 10 MM11

definitely NONE of margin indistinct due to tissue invasion — definitely some of margin indistinct due to tissue invasion

- Confidence that at least a small portion of the margin is spiculated

Current 0 1 2 3 4 5 6 7 8 9 10 MM12

definitely NOT spiculated — definitely spiculated

- Percentage of the margin that is . . . (total should add to up 100%)

	Current	Prior	
A. Clearly circumscribed	_____%	_____%	MM13A
B. Obscured by glandular tissue	_____%	_____%	MM13B
C. Indistinct due to tissue invasion	_____%	_____%	MM13C
D. Spiculated	_____%	_____%	MM13D
Total	100%	100%	

- Degree of microlobulation

Current 0 1 2 3 4 5 6 7 8 9 10 MM14
NONE extensive

- Confidence that the mass is a skin lesion

Current 0 1 2 3 4 5 6 7 8 9 10 MM15
definitely NOT a skin lesion definitely a skin lesion

- Confidence that the mass is an intramammary node

Current 0 1 2 3 4 5 6 7 8 9 10 MM16
definitely NOT an intramammary node definitely an intramammary node

- Confidence regarding the presence of related architectural distortion

Current 0 1 2 3 4 5 6 7 8 9 10 MM17
definitely NOT present definitely present

- Confidence regarding the presence of worrisome calcifications within the mass

Current 0 1 2 3 4 5 6 7 8 9 10 MM18

definitely NOT present — definitely present

- Confidence regarding the presence of benign calcifications within the mass

Current 0 1 2 3 4 5 6 7 8 9 10 MM19

definitely NOT present — definitely present

- Is there an ultrasound for the current images?
  - O Yes
  - O No MMUL

If no ultrasound, skip to bottom of page 5, Overall Diagnostic Judgment.

Questions relating to Ultrasound for the Current Images

- Appearance of the mass wall (ultrasound)
  - O well circumscribed
  - O indistinct
  - O irregular MMUL1

- Contents of the mass (ultrasound)
  - O solid
  - O indeterminate
  - O cystic MMUL2

- Response of the posterior wall of the mass (ultrasound)
  - O enhancement
  - O iso-echoic
  - O shadowing MMUL3

- Shape of the mass (ultrasound)
  - O round
  - O ellipsoid
  - O irregular MMUL4

Overall Diagnostic Judgment

- **Benign vs. Malignant**
  Rate the likelihood (as the number of chances in 100) that the finding is indicative of malignancy:

  Rating (0 to 100) _____ **MMRA**

  where: 0 = certainly benign or normal

  100 = certainly malignant

Appendix C (continued)

Illustrative Case as Read by Four Mammographers, With Three Rating Scales

Case No. 2

Reader No. R1

Panel Member: _____

DICTATED REPORT

Exam: #189 8/11/94 Mammography Bilateral

Bilateral Mammogram. The breast parenchyma is mild to moderately dense and nodular as seen on examination of December 8, 1992. Since the prior study, there has been interval development of a partially circumscribed 6 mm nodule in the left upper mid breast.
Left breast ultrasound: Examination of the 12 o'clock axis shows a 7 mm circumscribed relatively anechoic lesion with increased through transmission consistent with a simple cyst.

IMPRESSION: Simple cyst left mid upper breast.

a. Degree to which report is expressed in clear and standard terminology.

0 1 2 3 4 5 6 7 8 9 10
NOT in clear standard terms — In clear standard terms

b. Adequacy of detail in the report.

0 1 2 3 4 5 6 7 8 9 10
Too little detail — Adequate detail — Too much detail

c. Fluency of text.

0 1 2 3 4 5 6 7 8 9 10
Stilted text — Fluent text

AUTOMATED REPORT

Exam: #189 8/11/94 Mammography Bilateral

The present examination is compared to a prior mammogram of 12/8/92. The breast is largely fibroglandular.

There is a new 6 mm oval mass at 12 o'clock in the middle region of the left breast. About half of the margin appears to be obscured by glandular tissue. A large portion of the margin is clearly circumscribed. Microlobulation is present. The mass may be an intramammary node.

An ultrasound was performed. The mass appears ellipsoid with indeterminate contents. The mass wall is well-circumscribed. The posterior wall of the mass displays enhancement.

IMPRESSION: There is a new 6 mm oval mass in the left breast. This lesion is almost certainly benign. The probability of malignancy is less than .05.

a. Degree to which report is expressed in clear and standard terminology.

0 1 2 3 4 5 6 7 8 9 10
NOT in clear standard terms — In clear standard terms

b. Adequacy of detail in the report.

0 1 2 3 4 5 6 7 8 9 10
Too little detail — Adequate detail — Too much detail

c. Fluency of text.

0 1 2 3 4 5 6 7 8 9 10
Stilted text — Fluent text

Case No. 2

Reader No. R2

Panel Member: _____

DICTATED REPORT

Exam: #189 8/11/94 Mammography Bilateral

Comparison is made to a prior study dated 12/8/92. The breast parenchyma is moderately dense. In the left upper, slightly outer breast, a 7 x 6 mm. low density partially circumscribed and partially obscured nodule is noted. This is new since the prior study. Further evaluation with breast ultrasound demonstrates a corresponding 6 mm. simple cyst at 1:00.

IMPRESSION:
1. Moderately dense parenchyma.
2. Circumscribed 7 mm. nodule left upper, slightly outer breast, with a corresponding simple cyst on breast ultrasound.
3. No suspicious findings.

a. Degree to which report is expressed in clear and standard terminology.

0 1 2 3 4 5 6 7 8 9 10
NOT in clear standard terms — In clear standard terms

b. Adequacy of detail in the report.

0 1 2 3 4 5 6 7 8 9 10
Too little detail — Adequate detail — Too much detail

c. Fluency of text.

0 1 2 3 4 5 6 7 8 9 10
Stilted text — Fluent text

AUTOMATED REPORT

Exam: #189 8/11/94 Mammography Bilateral

The present examination is compared to a prior mammogram of 12/8/92. The breast is largely fat.

There is a new 8 mm oval mass at 12 o'clock in the middle region of the left breast. A small portion of the margin appears to be obscured by glandular tissue. A large portion of the margin is clearly circumscribed.

An ultrasound was performed. The mass appears ellipsoid with cystic contents. The mass wall is well-circumscribed. The posterior wall of the mass displays enhancement.

IMPRESSION: There is a new 8 mm oval mass in the left breast. This lesion is almost certainly benign. The probability of malignancy is less than .05.

a. Degree to which report is expressed in clear and standard terminology.

0 1 2 3 4 5 6 7 8 9 10
NOT in clear standard terms — In clear standard terms

b. Adequacy of detail in the report.

0 1 2 3 4 5 6 7 8 9 10
Too little detail — Adequate detail — Too much detail

c. Fluency of text.

0 1 2 3 4 5 6 7 8 9 10
Stilted text — Fluent text

Case No. 2

Reader No. R3

Panel Member: _____

DICTATED REPORT

Exam: #189 8/11/94 Mammography Bilateral

Examination dated 8/11/94. Bilateral Mammography: The breast parenchyma is mildly dense. In the interval since 12/8/92 a 5 mm low density smoothly marginated mass has developed in the 12 o'clock region of the left breast. This mass has no calcifications and 80 percent of the margin is well circumscribed. An ultrasound shows that this is oval hypoechoic and has the appearance of a cyst with debris. An aspiration is recommendation for further evaluation.

IMPRESSION: Guided aspiration recommended for 5 mm hypoechoic mass in the left breast.

a. Degree to which report is expressed in clear and standard terminology.

0 1 2 3 4 5 6 7 8 9 10
NOT in clear standard terms — In clear standard terms

b. Adequacy of detail in the report.

0 1 2 3 4 5 6 7 8 9 10
Too little detail — Adequate detail — Too much detail

c. Fluency of text.

0 1 2 3 4 5 6 7 8 9 10
Stilted text — Fluent text

AUTOMATED REPORT

Exam: #189 8/11/94 Mammography Bilateral

The present examination is compared to a prior mammogram of 12/8/92. The breast is almost entirely fat.

There is a new 10 mm oval mass at 12 o'clock in the middle region of the left breast. A small portion of the margin appears to be obscured by glandular tissue. A large portion of the margin is clearly circumscribed.

An ultrasound was performed. The mass appears ellipsoid with indeterminate contents. The mass wall is well-circumscribed. The posterior wall of the mass displays enhancement.

IMPRESSION: There is a new 10 mm oval mass in the left breast. This lesion is almost certainly benign. The probability of malignancy is less than .05.

a. Degree to which report is expressed in clear and standard terminology.

0 1 2 3 4 5 6 7 8 9 10
NOT in clear standard terms — In clear standard terms

b. Adequacy of detail in the report.

0 1 2 3 4 5 6 7 8 9 10
Too little detail — Adequate detail — Too much detail

c. Fluency of text.

0 1 2 3 4 5 6 7 8 9 10
Stilted text — Fluent text

Case No. 2

Reader No. R4

Panel Member: _____

DICTATED REPORT

Exam: #189 8/11/94 Mammography Bilateral

Bilateral mammogram is compared with an earlier study of 12/8/92. Breast parenchyma is moderately dense and then slightly nodular. A 6 mm. low density nodule in the left upper mid to slightly outer breast is demonstrated. Further evaluation with ultrasound was performed. Sonographic evaluation of the left breast in the 12 o'clock position revealed a 6 mm. oval nodule with increased through transmission; this most likely represents a benign cyst.

IMPRESSION: Moderately dense, slightly nodular breasts with new circumscribed nodule, left upper mid-outer breast, compatible with a small cyst. One-year follow-up examination is recommended.

a. Degree to which report is expressed in clear and standard terminology.

0 (NOT in clear standard terms) 1 2 3 4 5 6 7 8 9 10 (In clear standard terms)

b. Adequacy of detail in the report.

0 (Too little detail) 1 2 3 4 5 (Adequate detail) 6 7 8 9 10 (Too much detail)

c. Fluency of text.

0 (Stilted text) 1 2 3 4 5 6 7 8 9 10 (Fluent text)

AUTOMATED REPORT

Exam: #189 8/11/94 Mammography Bilateral

The present examination is compared to a prior mammogram of 12/8/92. The breast is almost entirely fat.

There is a new 8 mm oval mass at 12 o'clock in the middle region of the left breast. A large portion of the margin appears to be obscured by glandular tissue. Some of the margin is clearly circumscribed.

An ultrasound was performed. The mass appears ellipsoid with cystic contents. The mass wall is well-circumscribed. The posterior wall of the mass displays enhancement.

IMPRESSION: There is a new 8 mm oval mass in the left breast. This lesion is almost certainly benign. The probability of malignancy is less than .05.

a. Degree to which report is expressed in clear and standard terminology.

0 (NOT in clear standard terms) 1 2 3 4 5 6 7 8 9 10 (In clear standard terms)

b. Adequacy of detail in the report.

0 (Too little detail) 1 2 3 4 5 (Adequate detail) 6 7 8 9 10 (Too much detail)

c. Fluency of text.

0 (Stilted text) 1 2 3 4 5 6 7 8 9 10 (Fluent text)

Appendix C (continued)

Rating Scales for Report Consistency Across Mammographers

Panel Member: _____

Case	DICTATED REPORT	AUTOMATED REPORT
1	d. Consistency of terminology across the four reports. 0 1 2 3 4 5 6 7 8 9 10 Very inconsistent — Very consistent e. Consistency of content across the four reports. 0 1 2 3 4 5 6 7 8 9 10 Very inconsistent — Very consistent	d. Consistency of terminology across the four reports. 0 1 2 3 4 5 6 7 8 9 10 Very inconsistent — Very consistent e. Consistency of content across the four reports. 0 1 2 3 4 5 6 7 8 9 10 Very inconsistent — Very consistent
2	d. Consistency of terminology across the four reports. 0 1 2 3 4 5 6 7 8 9 10 Very inconsistent — Very consistent e. Consistency of content across the four reports. 0 1 2 3 4 5 6 7 8 9 10 Very inconsistent — Very consistent	d. Consistency of terminology across the four reports. 0 1 2 3 4 5 6 7 8 9 10 Very inconsistent — Very consistent e. Consistency of content across the four reports. 0 1 2 3 4 5 6 7 8 9 10 Very inconsistent — Very consistent
3	d. Consistency of terminology across the four reports. 0 1 2 3 4 5 6 7 8 9 10 Very inconsistent — Very consistent e. Consistency of content across the four reports. 0 1 2 3 4 5 6 7 8 9 10 Very inconsistent — Very consistent	d. Consistency of terminology across the four reports. 0 1 2 3 4 5 6 7 8 9 10 Very inconsistent — Very consistent e. Consistency of content across the four reports. 0 1 2 3 4 5 6 7 8 9 10 Very inconsistent — Very consistent
4	d. Consistency of terminology across the four reports. 0 1 2 3 4 5 6 7 8 9 10 Very inconsistent — Very consistent e. Consistency of content across the four reports. 0 1 2 3 4 5 6 7 8 9 10 Very inconsistent — Very consistent	d. Consistency of terminology across the four reports. 0 1 2 3 4 5 6 7 8 9 10 Very inconsistent — Very consistent e. Consistency of content across the four reports. 0 1 2 3 4 5 6 7 8 9 10 Very inconsistent — Very consistent

Panel Member: _____

Case	DICTATED REPORT	AUTOMATED REPORT
5	d. Consistency of terminology across the four reports. 0 1 2 3 4 5 6 7 8 9 10 Very inconsistent — Very consistent e. Consistency of content across the four reports. 0 1 2 3 4 5 6 7 8 9 10 Very inconsistent — Very consistent	d. Consistency of terminology across the four reports. 0 1 2 3 4 5 6 7 8 9 10 Very inconsistent — Very consistent e. Consistency of content across the four reports. 0 1 2 3 4 5 6 7 8 9 10 Very inconsistent — Very consistent
6	d. Consistency of terminology across the four reports. 0 1 2 3 4 5 6 7 8 9 10 Very inconsistent — Very consistent e. Consistency of content across the four reports. 0 1 2 3 4 5 6 7 8 9 10 Very inconsistent — Very consistent	d. Consistency of terminology across the four reports. 0 1 2 3 4 5 6 7 8 9 10 Very inconsistent — Very consistent e. Consistency of content across the four reports. 0 1 2 3 4 5 6 7 8 9 10 Very inconsistent — Very consistent
7	d. Consistency of terminology across the four reports. 0 1 2 3 4 5 6 7 8 9 10 Very inconsistent — Very consistent e. Consistency of content across the four reports. 0 1 2 3 4 5 6 7 8 9 10 Very inconsistent — Very consistent	d. Consistency of terminology across the four reports. 0 1 2 3 4 5 6 7 8 9 10 Very inconsistent — Very consistent e. Consistency of content across the four reports. 0 1 2 3 4 5 6 7 8 9 10 Very inconsistent — Very consistent
8	d. Consistency of terminology across the four reports. 0 1 2 3 4 5 6 7 8 9 10 Very inconsistent — Very consistent e. Consistency of content across the four reports. 0 1 2 3 4 5 6 7 8 9 10 Very inconsistent — Very consistent	d. Consistency of terminology across the four reports. 0 1 2 3 4 5 6 7 8 9 10 Very inconsistent — Very consistent e. Consistency of content across the four reports. 0 1 2 3 4 5 6 7 8 9 10 Very inconsistent — Very consistent

PUBLICATIONS, ABSTRACTS, AND PERSONNEL

Project Title:	Increasing the Accuracy of Mammogram Interpretation
Principal Investigator:	John A. Swets, Ph.D.
Contract No.	DAMD17-94-C-4082
Contractor:	Bolt Beranek and Newman Inc. Systems and Technologies Division Now: BBN Technologies A Unit of GTE Internetworking 10 Moulton Street Cambridge, MA 02138
Subcontractor:	Brigham and Women's Hospital Radiology Department 25 Francis Street Boston,MA 02115
Principal Investigator:	Steven E. Seltzer, M.D.
Collaborating Institution:	Harvard Pilgrim Health Care Radiology Department 2 Essex Center Drive Peabody, MA 01960
Principal Investigator:	William J. Otto, Jr., M.D.
Term of Contract:	15 September 1994 - 14 February 1999 (research ended 14 September 1998)
Date:	14 October 1998

Publications

Getty, D. J. Assisting the radiologist to greater accuracy. In SPIE Medical Imaging 1996: Physics of Medical Imaging, R. L. Van Metter and J. Bentel (Eds.), 2708, 2-15.

Swets, J. A. Separating discrimination and decision in detection, recognition, and matters of life and death. In Osherson, D. (series editor), Sternberg, S. and Scarborough, D., (Vol. Eds.) An invitation to cognitive science: Vol. 4. Methods, models, and conceptual issues. Cambridge, MA: MIT Press, 1998, 635-702.

Swets, J. A. Enhancing diagnostic decisions. In Hoffman, R. R., Sherrick, M. F., and Warm, J. S. (Eds.), Viewing psychology as a whole. Washington, DC: American Psychological Association, 1998, 559-577.

Swets, J. A. Enhancing diagnostic decisions. In Connolly, T., Arkes, H. R., and Hammond, K. R. (Eds.), Judgment and Decision Making: An Interdisciplinary Reader (2nd Edition). Cambridge: Cambridge University Press (in press).

Related Publications (As defined by the American Institute of Biological Sciences, 24 April 1998, AIBS #0591.)

Swets, J. A. Signal Detection Theory and ROC Analysis in Psychology and Diagnostics, Mahwah, NJ: Erlbaum, 1996.

Seltzer, S. E., Getty, D. J., Tempany, C.M.C., Pickett, R. M., Schnall, M. D., McNeil, B. J., and Swets, J. A. Staging prostate cancer with MR imaging: A combined radiologist-computer system. Radiology, 1997; 202(1):219-226.

Getty, D. J., Seltzer, S. E., Tempany, C.M.C., Pickett, R. M., Swets, J. A., and McNeil, B. J. Prostate cancer: Relative effects of demographic, clinical, histologic, and MR imaging variables on the accuracy of staging. Radiology, 1997; 204(2):471-479.

Meeting Presentations

Getty, D. J. Assisting the radiologist to greater accuracy. Keynote Address, SPIE, Newport Beach, CA, 11-13 February 1996.

Swets, J. A. Enhancing diagnostic decisions. In Psychology Beyond the Threshold, University of Cincinnati, 15-18 May 1995.

Swets, J. A. Decision aids for radiologists. American Association for the Advancement of Science, Baltimore, MD, 10-14 February 1996.

Swets, J. A. Discussant, Signal Detection Theory Across the Discipline: It's Not Just d'. Invited Symposium, American Psychological Society, Washington, D.C., 22-26 May 1997.

Swets, J. A. From esoteric psychophysics to wordly diagnostics. In Psychology Works: From Basic Research to Better Mousetraps. Presidential Symposium, American Psychological Society, Washington, D.C., 21-24 May 1998.

Meeting Abstracts

Following are the extended and lay/public abstracts presented with a poster at the Breast Cancer Research Program: An Era of Hope, Renaissance Hotel, Washington, D.C., 31 October - 4 November 1997.

INCREASING THE ACCURACY OF MAMMOGRAM INTERPRETATION

Computer-Based System for Decision Support and Automated Reporting

**John A. Swets, David J. Getty, Ronald M. Pickett,
Steven E. Seltzer, and William J. Otto, Jr.**

BBN Technologies, a business unit of GTE Internetworking (Cambridge MA),
Brigham and Women's Hospital (Boston MA),
and Harvard Pilgrim Health Care (Boston MA)

The aims of the study are to develop a computer-based system that will aid the radiologist in interpreting mammograms, automatically provide a standardized report of mammogram findings to a referring physician, and construct a database of results to help assure the quality of the interpretive process.

In a completed system, the radiologist will assign a rating-scale value to each of a set of perceptual features that have been statistically determined to be diagnostically relevant and comprehensive. Spoken scale values will be recognized by the system, merged optimally (in terms of their predictive weights and intercorrelations) to yield an estimate of the probability of malignancy, and analyzed interactively to generate automatically a prose report using the lexicon of the American College of Radiology. A database organized about the perceptual features will help to resolve differences in dual readings, to construct tutorial materials tailored to individual radiologists, and to adjust thresholds for recommendations of follow-on imaging or therapy.

Five mammographers at Brigham and Women's Hospital (BWH) have assigned ratings to a large set of perceptual features for 200 proven BWH cases both to determine a necessary and sufficient set of features and to "train" a statistical prediction rule to estimate the probability of malignancy as based on feature ratings. Five radiologists at Harvard Pilgrim Health Care (HPHC) will assign ratings to features in a set of 150 proven HPHC cases to "test" the statistical prediction rule. The mammograms thus interpreted at HPHC -- with the aid both of the list of perceptual features and the probability estimate of malignancy -- will be compared to baseline interpretations of the same cases obtained there earlier from the same radiologists, in order to determine the gain in accuracy provided by the computer-based aid. Reports of mammogram findings generated automatically by the system for selected cases will be assessed by a group of referring physicians and surgeons relative to reports dictated for those cases in the usual way.

This project builds on work done previously in the BBN laboratory to increase accuracy and extends it into the clinic of both a referral center (BWH) and screening site (HPHC). It extends previous work to incorporate the reporting process.

The statistical prediction rule developed earlier led to significant accuracy enhancements and has now been refined in certain ways, principally by considering changes in the perceptual features from prior to current mammograms. The linear-

discriminant analysis used earlier as the technique to create the rule was replaced by the logistic-regression technique.

We made assessments of accuracy by ROC analysis (relative, or receiver, operating characteristic) to obtain an index of accuracy that is unaffected by an observer's decision threshold and by the relative frequencies (prior probabilities) of malignant and non-malignant cases in the test set. The ROC analysis was made directly from the estimates of probability of malignancy made by the statistical prediction rule and also from estimates made by radiologist observers after receiving the rule's estimate as an advisory.

Cases were obtained retrospectively at the two clinical sites and were selected to represent malignancies, benign lesions, and "suspicious" cases that were determined subsequently to be "normal." Images taken at two different times were included -- the images first deemed suspicious and the images of the last preceding examination.

The radiologist observers are representative of the referral and community-hospital settings, respectively -- the former being more highly specialized in mammography. The statistical prediction rule, or decision aid, is thus as effective as specialists can make it, but, we think, still suitable for the different case mixes of various screening settings.

The checklist of all diagnostically important perceptual features is an aid to the radiologist in making a complete assessment of image information, so not to be lulled by a premature "satisfaction of search" when a few dominant features appear. Following the checklist, however, may take additional time and the radiologist may choose to use it only for difficult cases. In the envisioned practical computer system, desired cases can be selected for system application in a seamless way, interwoven with cases not selected; with computer-based speech recognition, the microphone usually used for dictation controls the use of the decision aids. Additional motivation for using the decision system, beyond enhancements of accuracy, include the automated report as well as a rich database of diagnostic findings and treatment outcomes on all cases to help assure quality in several ways. Overall, we expect the cost-benefit tradeoff to favor system use.

Keywords: **Mammography Interpretation, Perceptual-Feature Analysis, Computer-Assisted Diagnosis, Automated Reports, Increased Accuracy and Standardization**

This work was supported by the U.S. Army Medical Research and Materiel Command under DAMD17-94-C-4082.

INCREASING THE ACCURACY OF MAMMOGRAM INTERPRETATION

Dr. John A. Swets, Dr. David J. Getty, Dr. Ronald M. Pickett, Dr. Steven E. Seltzer, and Dr. William J. Otto, Jr.

BBN Corporation, Brigham and Women's Hospital,
and Harvard Community Health Plan
Cambridge, MA 02138

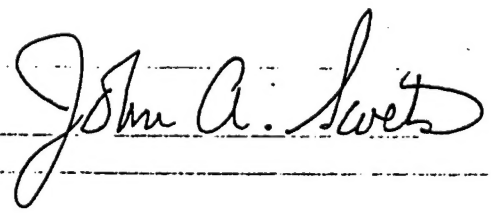
A computer-based system under development is intended to increase the radiologist's accuracy in interpreting mammograms. For each patient, the computer leads the radiologist through all of the dozen or so perceptual features of a mammogram that are important to diagnosis (for example, indistinctness of the border of a visible mass, indicating a possible invasion of a tumor into surrounding tissue). For each feature, the computer solicits from the radiologist a quantitative value (for example, a rating on a ten-point scale of the degree to which a mass's border is indistinct).

The several scale values are then combined by the computer in an optimal manner to produce for the radiologist an estimate of the probability of malignancy. This probability estimate has been shown to be more accurate than the radiologist's unaided estimate and it offers a significant increase in the accuracy of diagnosis.

The radiologist's several scale values are also translated automatically into a standardized prose report for the referring physician and surgeon. This report is linguistically and structurally apt and uses the lexicon recommended by the American College of Radiology.

Lastly, the scale values are stored with clinical data for each case (history, other findings, therapeutic outcomes) so that a database on all patients is available to help monitor and assure the quality of the radiological process, with reference both to the individual radiologist and the medical center.

The advantages of this research to the field and the public are an increase in diagnostic accuracy, the controlled standardization of the radiologist's report of findings, and a database centered on perceptual features of the mammogram to assist the assurance of quality in the diagnostic process.



Information Submitted to U.S. Public Health Service's Office on Women's Health for September 1966 Congressional Briefing

Project Abstract (Emphasize source of technology, how being applied, scope of project)

We are refining and testing in the clinic a computer-based decision-aiding system for mammogram interpretation. For each of the most relevant perceptual features, the radiologist speaks a scale value or rating. These spoken ratings are recognized by the computer and are optimally and automatically converted into (i) a probability of malignancy and (ii) a standardized prose report of findings. Twelve radiologists, from both screening and referral settings, will read large sets of proven cases. Focus groups of physicians will help explore the role of the system in quality assurance, continuing education, and networked communications.

Potential Benefits to Breast Imaging

The feature-based decision-aiding system was earlier demonstrated to improve diagnostic accuracy very substantially. Such a system should increase the usefulness and validity of prose reports to the clinician and surgeon. Quantitative treatment in terms of probability of malignancy will refine decision thresholds for the various levels of treatment - from routine follow-up, to accelerated follow-up, to other types of imaging examinations, to biopsy - offering enhancement both of patient care and utilization of diagnostic resources.

Current Status/Results Achieved

An appropriate set of proven cases has been defined and culled from files of the Brigham and Women's Hospital (BWH) and another is being identified at the Harvard Community Health Plan (HCHP). In the fall of 1996, five BWH radiologists will do baseline readings of their cases and then re-read them with a recently completed questionnaire to assign the perceptual-feature ratings that will train the decision aid. A speech-recognizer was adapted to accept feature ratings. The automated report writer is under active development to provide linguistically and clinically sensible reports.

When is clinical application projected?

HCHP radiologists will provide final testing in the early spring of 1997. Focus groups of radiologists and clinicians will be convened then to explore clinical application.

Is there a suitable demonstration available for the Congressional Briefing? If so, in what form?

Ready soon is a computer system that elicits spoken ratings of perceptual features from radiologists as they view mammograms. Graphs on posters could show the improved accuracy of diagnosis that has been demonstrated. Likely to be ready in the spring of 1997 is the automated report writer that generates prose reports from spoken ratings.

Principal Investigator: John A. Swets, Ph.D.

PROJECT STAFF

John Swets (P.I.), Ph.D.	BBN Technologies
David J. Getty, (Co P.I.) Ph.D.	BBN Technologies
Ronald M. Pickett, Ph.D.	BBN Technologies
Marie Meteer, Ph.D.	BBN Technologies
Nancy Babiarz, BA	BBN Technologies
Robert Granville, BA	BBN Technologies
Barbara Freeman, BA	BBN Technologies
Steven E. Seltzer, (Co P.I.) M.D.	Brigham and Women's Hospital
Edward Chao, BS	Brigham and Women's Hospital
Lisa Beth Cronen, BS	Brigham and Women's Hospital
Sue Calder, BA	Brigham and Women's Hospital
Heather Gibbons, BA	Brigham and Women's Hospital
Dianne Johnson, BA	Brigham and Women's Hospital
Lisa Herrman, Ph.D.	Brigham and Women's Hospital
William J. Otto, Jr., M.D.	Harvard Pilgrim Health Care
Carl J. D'Orsi, M.D.	University of Massachusetts Medical Center

RADIOLOGIST READERS

BWH:
Christine Denison, M.D
Pamelo DiPiro, M.D.
Thomas Frenna, M.D.
Jack Meyer, M.D.
Darrell Smith, M.D.

HPHC:
Stephen Barrand, M.D.
Jeffrey Melamed, M.D.
Jean O'Brien, M.D.
William Otto, Jr., M.D.
Philip Thomason, M.D.